



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁷ : C12Q 1/68		A2	(11) International Publication Number: WO 00/65088 (43) International Publication Date: 2 November 2000 (02.11.00)
(21) International Application Number: PCT/EP00/03636 (22) International Filing Date: 20 April 2000 (20.04.00)		(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).	
(30) Priority Data: 99303215.0 26 April 1999 (26.04.99) EP		Published Without international search report and to be republished upon receipt of that report.	
(71) Applicant (<i>for all designated States except US</i>): AMERSHAM PHARMACIA BIOTECH AB [SE/SE]; Bjorkgatan 30, S-751 84 Uppsala (SE).			
(72) Inventors; and (75) Inventors/Applicants (<i>for US only</i>): ULFENDAHL, Per-Johan [SE/SE]; Rapphovsvagen 10B, S-756 53 Uppsala (SE). WONG, Kin-Chun [SE/SE]; Ursviksvagen 2B, S-172 36 Sundbyberg (SE).			
(74) Agent: ROLLINS, Anthony, John; Nycomed Amersham plc, Amersham Laboratories, White Lion Road, Amersham, Bucks HP7 9LL (GB).			

(54) Title: PRIMERS FOR IDENTIFYING TYPING OR CLASSIFYING NUCLEIC ACIDS

(57) Abstract

A method is described for identifying a rather small set of extendible primers for use in the identification, typing or classification of a nucleic acid of known sequence having known polymorphisms. A matrix of primers and pairs of primer extensions is prepared and subjected to analysis by a set covering problem algorithm, e.g. a greedy algorithm or one which involves a Lagrangian relaxation heuristic. Sets of primers are described for use in the identification, classification or typing of an organism, allele or gene selected from class I HLA, class 2 HLA and 16S rRNA.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	ES	Spain	LS	Lesotho	SI	Slovenia
AM	Armenia	FI	Finland	LT	Lithuania	SK	Slovakia
AT	Austria	FR	France	LU	Luxembourg	SN	Senegal
AU	Australia	GA	Gabon	LV	Larvia	SZ	Swaziland
AZ	Azerbaijan	GB	United Kingdom	MC	Monaco	TD	Chad
BA	Bosnia and Herzegovina	GE	Georgia	MD	Republic of Moldova	TG	Togo
BB	Barbados	GH	Ghana	MG	Madagascar	TJ	Tajikistan
BE	Belgium	GN	Guinea	MK	The former Yugoslav Republic of Macedonia	TM	Turkmenistan
BF	Burkina Faso	GR	Greece	ML	Mali	TR	Turkey
BG	Bulgaria	HU	Hungary	MN	Mongolia	TT	Trinidad and Tobago
BJ	Benin	IE	Ireland	MR	Mauritania	UA	Ukraine
BR	Brazil	IL	Israel	MW	Malawi	UG	Uganda
BY	Belarus	IS	Iceland	MX	Mexico	US	United States of America
CA	Canada	IT	Italy	NE	Niger	UZ	Uzbekistan
CF	Central African Republic	JP	Japan	NL	Netherlands	VN	Viet Nam
CG	Congo	KE	Kenya	NO	Norway	YU	Yugoslavia
CH	Switzerland	KG	Kyrgyzstan	NZ	New Zealand	ZW	Zimbabwe
CI	Côte d'Ivoire	KP	Democratic People's Republic of Korea	PL	Poland		
CM	Cameroon	KR	Republic of Korea	PT	Portugal		
CN	China	KZ	Kazakhstan	RO	Romania		
CU	Cuba	LC	Saint Lucia	RU	Russian Federation		
CZ	Czech Republic	LI	Liechtenstein	SD	Sudan		
DE	Germany	LK	Sri Lanka	SE	Sweden		
DK	Denmark	LR	Liberia	SG	Singapore		
EE	Estonia						

PRIMERS FOR IDENTIFYING TYPING OR CLASSIFYING NUCLEIC ACIDS

5

DNA-sequence analysis is rapidly becoming a standard tool in modern, molecular biology research. Examples of applications include: Sequencing of unknown DNA-sequences, Identifying novel genes in stretches of sequenced DNA, Predicting protein-sequence and -structure 10 from DNA-sequence alone and Identification of known gene-variations (sometimes called "typing a gene").

Typing of a gene could be crucial in some applications. For instance, organ-donation requires that the "immunological signature" of the donor matches that of the receiver. This "signature" is mediated by the 15 *Human Leucocyte Antigen (HLA) complexes* (also known as *Major Histocompatibility Complex, MHC*) on the cell surface, and the corresponding genes are among the most varied in the human genome. Considering the importance of organ donation, the shortage of organ-donors and the fact that an organ cannot be stored for any longer time-periods, a rapid and accurate typing of the HLA-genes is required in order 20 to make most use of the organs available for transplantations.

Another application where a rapid and accurate identification of a gene is desired is when trying to identify unknown bacteria. A rapid identification of the bacteria causing the illness of a patient makes it 25 possible to administer the correct medication early in the treatment of the disease, thus reducing the discomfort for the patient. Since every self-replicating organism so far studied use ribosomes when translating mRNA to proteins, analysis of one of the genes coding for the ribosome, for instance the 16S rRNA in the case of prokaryotes, could be used to identify 30 the organism in question.

There are several ways in which a gene can be identified, with the conceptually easiest being to sequence the entire gene and then

looking at the result. The main drawback is that this approach is time-consuming, and not easily scaled up using conventional methodology. A new method, *Arrayed Primer EXtension (APEX)*, lacks this drawback. APEX works by immobilising a large number of primers to a solid surface, thus creating a DNA-chip. These primers are constructed to be consecutively overlapping over the entire gene of interest, so that every base in the gene will have a primer to its 5'-end. By adding fluorescently labelled dideoxynucleotides, the primers will then be extended by one nucleotide using the sample DNA as template. It will thus be easy to check which nucleotide was incorporated, which in turn tells you the entire sequence of the sample DNA.

Since some genes, like the HLA and 16S rRNA, have a large number of known variations, a prohibitively large number of primers have to be created in order to probe for all possible combinations of variant positions in the gene. Thus the array primer extension method APEX for resequencing would need more than 16,000 primers if all DQB alleles would be sequenced from a 500 bp long PCR fragment. If all DQB alleles in pairs should be combined the number of primers might be even higher which would be the situation for a heterozygote found in most individuals.

But this might not be necessary, if some variations always or never occur together. This needs to be studied though, and a way found to determine the least number of primers (and what their sequences are) required for unambiguously identifying those genes.

An object of this invention is to find and implement an efficient algorithm capable of doing just that. The algorithm should preferably also take into account the melting points of the primers, so that the extension reaction can take place under optimal conditions for all of the primers on the chip. It should also minimise the number of "self-extended" primers, i.e. primers that can extend themselves without any sample DNA. This algorithm is then to be tested and evaluated on the HLA and 16S rRNA-genes. HLA is chosen partly because of the importance of rapid typing of these genes, leading to the fact that there are many other methods to

which APEX can be compared. It is also because the HLA-genes are "easy" to work with, since they rarely contain any insertions or deletions. These kinds of variations in the gene could potentially create problems when designing primers for APEX. The 16S rRNA, on the other hand, 5 contains insertions and deletions and can thus be used to see if the algorithm can handle such variations.

The invention provides a method of identifying a set of extendible primers for use in the identification, typing or classification of a nucleic acid of known sequence having known polymorphisms wherein:

- 10 i) all possible nucleotide sequences of a chosen length of the nucleic acid are identified and their corresponding extendible primers,
- ii) at least one extendible primer is removed from the set wherein the at least one primer removed identifies a segment of the nucleic acid identified by at least one other primer.

15 Preferably the method includes between step i) and ii):

- ia) potential extensions for each primer are identified with respect to each nucleotide sequence,
- ib) for each extendible primer the identified potential extensions are compared to determine which pairs of sequences can be discriminated

20 by the primer.

Preferably a matrix of primers and pairs of primer extensions is prepared in binary form and is subjected to analysis by a set covering problem (SCP) algorithm as described in more detail below.

The invention also includes a set of extendible primers, for 25 use in the identification, typing or classification of a nucleic acid of known sequence having known polymorphisms, identified by the method as defined. Preferably the primers are attached by 5'-ends to a surface of a support on which they are presented in the form of an array.

In another aspect, the invention provides a set of extendible 30 primers, for use in the identification, typing or classification of a human leucocyte antigen (HLA) gene as indicated, the set comprising about the

number of primers indicated and being capable of distinguishing about the number of alleles indicated:

	HLA gene	Number of Alleles	Number of Primers
Class I	HLA-A	91	172
	HLA-B	200	<1000
	HLA-C	47	94
Class II	DPA-1	11	26
	DRB-1	74	130
	DQA-1	17	130
	DQB-1	34	84
	DRB-1	192	<1000
	DRB345	35	94

5 In another aspect, the invention provides a set of extendible primers, for use in the identification, typing or classification of 16S rRNA, wherein the set comprises about 210 primers and is capable of distinguishing at least about 1207 different sequences.

10 In these aspects of the invention, the approximate number of primers is indicated. As indicated below, it may be possible by the use of the algorithms exemplified or other algorithms to generate slightly smaller sets of primers capable of distinguishing the number of alleles or sequences indicated, and these sets are envisaged according to the invention. Of course, other primers may be present in addition to those indicated as essential, and may be useful for checking purposes. The 15 number of alleles or sequences indicated represents the approximate known number of polymorphisms or different sequences, and these will surely increase with time.

20 In another aspect the invention provides a method of identification, typing or classification of a nucleic acid of known sequence having known polymorphisms, by the use of the set of extendible primers as defined, which method comprises applying the nucleic acid or fragments thereof to the set of extendible primers under hybridisation conditions and effecting template-directed chain extension of extendible primers that have

formed hybrids. Preferably template-directed chain extension is effected using four different fluorescently labelled chain-terminating nucleotide analogues, and results are analysed by an imaging system such as total internal reflection fluorescence (TIRF) or scanning confocal microscopy.

5 The various steps of the method may be performed as described in the literature for the known APEX technique.

In another aspect the invention provides a kit for use in the identification, typing or characterisation of a nucleic acid of known sequence having known polymorphisms, comprising the set of extendible primers as defined.

10 In another aspect the invention provides an array of sets of extendible primers as defined, for the simultaneous identification, typing or classification of two or more different HLA genes.

15 With the present invention it has been realised that where a number of different alleles are to be identified, the total number of primers required to distinguish each of the alleles could be reduced as some primers would be common to all of the alleles, for example. Thus, with the present invention complete sets of primers for identification of each allele are identified and then the total number of primers in the combined sets is 20 reduced using predetermined rules.

Furthermore the present invention is based on the premise that as the primers are used to identify the presence or absence of a particular nucleotide sequence in any allele, the specific nucleotide that extends any particular primer is of less relevance than simply whether the 25 primer has been extended. Thus, the problem of reducing the overall number of primers is greatly simplified rendering the problem one suitable for treatment as a Set Covering Problem (SCP).

30 Embodiments of the present invention will now be described by way of example with reference to the accompanying drawings and examples, in which:

Figure 1 is a diagram of a signal matrix in accordance with the present invention;

Figure 2 is a diagram of the corresponding binary matrix for the signal matrix of Figure 1;

Figure 3 is a flow diagram of the steps for reducing the primer set in accordance with the present invention.

5 The following is an explanation to assist in an understanding of the principles underlying the manner in which the number of primers used in the identification of a plurality of sequences may be reduced.

10 Theoretically the number of primers required to identify k sequences grows as $O(k \cdot l)$, where l is the length of the sequences as each sequence requires l primers. However, the less the sequences differ from one another, the fewer primers are required as many of the primers required for identification of a first sequence may also be of use in identification of another sequence. This effect becomes more pronounced the greater the number of sequences to be identified and the greater the similarities.

15 Considering an initial set of n primers required in the identification of k sequences, a signal matrix of $k \times n$ can be constructed. Each element in the matrix represents the signal, if any, that is generated by a particular primer with respect to a particular sequence. The signal will either be one of the four nucleotides 'A', 'C', 'G', or 'T' or no signal '-'.
20 Figure 1 is an example of such a signal matrix where, for example, the signal generated by primer 2 with respect to sequence 3 is 'T'.

25 The signal matrix is then converted into a binary matrix that represents whether the signals for any particular primer differ with respect to different sequences. Thus, again with respect to primer 2, the same signal 'G' is generated for both sequences 1 and 2 but a different signal 'T' is generated with respect to sequence 3. The binary matrix is constructed by considering each column (each primer) of the signal matrix and comparing each signal in that column in turn. Thus, as shown in Figure 2, 30 the first row of the matrix represents a comparison of the signals for the first and second sequences, the second row represents a comparison of the signals for the first and third sequences and the third row represents a

comparison of the signals for the second and third sequences. Binary '0' represents the comparison revealing the same signal and binary '1' represents the comparison reveals different signals. In the case of primer 2, as mentioned earlier the signals for the first and second sequences are 5 the same ('0') whereas the signals for the first and third sequences are different ('1'). This conversion produces a matrix $m \times n$ where $m=(k(k-1))/2$. Hence, for large numbers of sequences, $2m$ grows approximately as the square of the number of sequences. Figure 2 shows the binary matrix for 10 the signal matrix of Figure 1.

As the primers are required to enable the differentiation of sequences from one another, the reduction of the signal matrix to a binary matrix, representing differences in the signals obtained for different sequences, distils that element of information necessary to enable a selection of the minimum number of primers necessary to identify the 15 individual sequences. From the binary matrix the least number of columns are selected such that each row contains at least one non-zero element. Thus, if one of the columns contained all '1's only that one column would be required. However, in the case of Figure 2, there is no single column containing all '1's and so two columns must be selected, for example 20 primers 1 and 2. Primers 1 and 2 together enable each of sequences 1, 2 and 3 to be differentiated and so the remaining primers are redundant.

Where large numbers of sequences and primers are involved, the binary matrix renders the data contained within that matrix suitable for mathematical analysis. Once the selection of the reduced number of 25 primers has been made, though, it is the signal matrix that is required during the use of the primers in the identification of the different sequences. Thus, the signal matrix is used to 'decode' the results of any analysis using the reduced number of primers.

In practice, large numbers of sequences and primers are 30 involved and the selection of a reduced set of primers cannot be performed by simple inspection of the binary matrix. For large numbers of primers, selection of a suitable reduced set of primers can be performed by treating

the selection as a Set Covering Problem (SCP). An SCP is an integer optimisation problem and is well known in fields such as airline crew scheduling, selecting manufacturing equipment and ingot mould selection in steel production. In such large scale problems that cannot be solved exactly (NP-hard), heuristics are used in order to generate a solution. As a SCP is NP-hard, global algorithms and algorithms that identify local optima are not very suitable on their own for a large scale SCP. They will simply require far too much computation, as they try to find a solution that can be proven to be at least locally optimal. For this reason heuristic methods are required instead. They do not claim to give even locally optimal solution, but are much faster.

Two known computational methods that have been found to be effective in identifying reduced sets of primers are the 'greedy' algorithm and Lagrangian relaxation algorithm.

15

Greedy Algorithm

The most simple heuristic is the *greedy* algorithm, where columns are added one at a time. The column to be added in each step is chosen so as to cover as many uncovered rows as possible (a row is covered if it has at least one non-zero element). In other words, if S_r is the set of columns already included in the solution at iteration r , and R_r is the set of rows with no non-zero elements at iteration r , column j_r is selected according to:

$$P_j = \sum_{i \in R_r} a_{ij}$$

$$j_r^* = \arg \min c_j / P_j \quad j \notin S_r$$

25

Equation 1

This continues until all rows are covered, or until no more columns exist which can cover any of the rows still uncovered. Instead of minimising the term c_j / P_j , other terms can be used. Example terms are c_j ,

$c_j / \log_2 P_j$ or $c_j / (P_j)^2$. Greedy algorithms of this type are described in "An Efficient Heuristic for Large Set Covering Problems", Vasko, Wilson, Naval Research Logistics Quarterly 1984, 31:163-171 the contents of which is incorporated herein by reference. The difference is in how much emphasis

5 to place on the cost of the column versus how many rows the column covers. It is shown, however, that this entire class of heuristics share the same worst case behaviour. If we denote the set of columns in the solution as S and the solution value as Z , then the worst case behaviour can be described as:

10

$$\frac{Z_{\text{heu}}}{Z_{\text{opt}}} \leq H(d)$$

Equation 2

where

$$Z = \sum_{j \in S} c_j x_j$$

$$H(d) = \sum_{j=1}^d \frac{1}{j}, \quad d = \max_j \sum_{i=1}^m a_{ij}$$

Equation 3

15

In other words, how much worse the heuristic solution is compared to the optimal solution is dependent on the maximum number of non-zero elements in the columns. The advantage is that this algorithm is fast, even though its time complexity is $O(m^2n)$ (there can be a maximum of m columns in the solution, i.e. the maximum number of iterations is m . For each iteration the matrix is traversed once to find the next column to be added). Altogether, we have that the time required to solve the problem in the worst case scenario will grow as the number of sequences to the power of five (four due to the number of rows, and one due to the number of columns). In the case of 16S rRNA (see later), where we have ~1000 sequences, the matrix will have ~500,000 rows. The number of primers

20

25

- 10 -

(columns) is in this case ~250,000.

Lagrangian relaxation

More sophisticated methods exist, which use other kinds of heuristics. One heuristic capable of generating the most optimal solutions is believed to be some kind of *Lagrangian relaxation* heuristic, where in each iteration the Lagrange multipliers for each column are used to calculate the Lagrangian cost for the columns. Such a Lagrangian relaxation heuristic is described in "A Heuristic Method for the Set Covering Problem", Capara et al Technical Report OR-95-8, Operations Research Group, University of Bologna 1995 the content of which is incorporated herein by reference. A near optimal vector of these costs is then calculated by a *subgradient* algorithm, before being used as input to a greedy algorithm. This is repeated until no improvements in the solution can be made.

In Lagrangian subgradient methods the *Lagrangian* of the original problem is considered instead of the original problem. In this case, the Lagrangian will be

$$L(u) = \min \sum_{j=1}^n c_j(u)x_j + \sum_{i=1}^m u_i$$

$$x_j = \begin{cases} 0 \\ 1 \end{cases}$$

20

Equation 4

where u_i is the Lagrangian multiplier for row i . $c_j(u)$ is the Lagrangian cost associated with column j , and is defined by

$$c_j(u) = c_j - \sum_{i=1}^m a_{ij}u_j$$

25

Equation 5

- 11 -

An optimal solution to Equation 4 is given by

$$x_j(u) = \begin{cases} 0 & \text{if } c_j(u) > 0 \\ 1 & \text{if } c_j(u) < 0 \\ 0 \text{ or } 1 & \text{if } c_j(u) = 0 \end{cases}$$

Equation 6

5 $L(u)$ can also be seen as an estimate of the lower bound for
 the solution, i.e. the sum of the costs for the columns in the optimal solution
 to the SCP will be $\geq L(u)$. The solution to the SCP can be found by finding
 an optimal multiplier vector u^* instead, but this will require much
 computation especially for a large SCP. But near-optimal multiplier vectors
 10 can be found within short time by using the *subgradient* vector $s(u)$, defined
 by

$$s_i(u) = 1 - \sum_{j=1}^n x_j(u), \quad i = 1 \dots m$$

Equation 7

15 u can be refined iteratively by using for example

$$u_i^{k+1} = \max \left\{ u_i^k + \lambda \frac{UB - L(u^k)}{\|s(u^k)\|^2} s_i(u^k), 0 \right\}$$

Equation 8

where $\lambda > 0$ is a step-size parameter and UB is an upper
 bound on the value of the solution. The initial u^0 can be defined arbitrarily.
 20 To solve the SCP, first a near-optimal multiplier vector u is found. This and
 Equation 6 is then used as a basis to form a feasible solution. The upper
 bound UB can then be updated to the value of this feasible solution (if it is
 better than the previous best solution), and a new near-optimal multiplier
 vector found and so on until convergence is reached.

Another alternative computational method that may be employed to solve such a SCP is 'surrogate relaxation' in which in each iteration a corresponding continuous problem is solved and made feasible before a sub-gradient algorithm is applied. Alternatively, genetic algorithms 5 may be employed in which the 'genome' consists of n bits, one bit for each of the columns.

It should also be borne in mind that as the SCP operates on the binary matrix which only represents differences in signals between sequences for the same primer, a primer in the selected reduced set may 10 generate a negative, '−', signal rather than a positive signal, A, C, G, T. To be sure that the sample does in fact contain a particular sequence it is essential to ensure that for each sequence at least one primer generates a positive signal. Furthermore, in practice redundancy is desirable as all reactions may not occur as intended. Therefore, the least number of 15 positive signals as well as the least number of differences in the signal pattern is preferably larger than one.

With reference to Figure 3, the following is a description of one method of selecting a reduced set of primers.

Firstly, all possible primers are selected (10) using the 20 standard APEX procedure to produce a first set of primers. During this selection a substring of the sequence to be analysed is used to construct one primer, then the substring is displaced by one base and another primer is constructed. This process is carried out from the start of the sequence until the entire sequence has been covered. Both strands of DNA are used 25 and this is repeated for all sequences. The primers should be long enough to be capable of discriminating between exact matches and mismatches involving one or two nucleotide pairs. Conveniently, the primers are 13bp long as this has been found to be sufficient to ensure the reaction, or longer to increase hybrid stability. However, to avoid steric hindrance on the chip 30 each primer may be 5'-tailed. In this example, twelve 'T's are added to the 5'-end of the primer so that the final length of the primers is 25bp.

Next all primers that are not suitable as primers are rejected

(12) and the rest is included in a primary primer set. Unsuitable primers are those where the three bases at the 3'-end are complementary to any substring of the primer. In some instances this can result in the primer being extended by a neighbouring primer and not the sample DNA as a template and for that reason such primers are considered unsuitable.

5 Also, any primers that would produce ambiguous signals are identified and rejected (14). A primer produces an ambiguous signal where it is not known which of the four bases is in the relevant position.

10 Each of the remaining primers in the primary set primer is then compared to each sequence in turn to determine whether the primer is extendible by each sequence and if the primer is extendible the base with which it would be extended is determined. A signal matrix of the primers with respect to each of the sequences is thus generated (16).

15 In order for a primer to be extended using the sample DNA as template, the three bases in the 3'-end of the primer must hybridise to the DNA. Otherwise the enzyme responsible for the extension will not be able to add a nucleotide to the primer. Of the rest of the primer (the poly-T tail excluded), at most two mismatches are allowed, otherwise the primer-DNA duplex is considered to be too unstable to be extended.

20 In ordinary PCR, all the bases must match in order for the primer to be extended. But then the temperature is raised to the melting point, T_m , of the primer in the extension step. In APEX, this reaction is carried out at 45°C, which is around 10°-20° below T_m of most primers. This means that the primers will hybridise to the DNA despite a few 25 mismatches, which is why two mismatches are allowed here.

In some cases a primer could hybridise to a sequence in more than one position, and sometimes a primer could hybridise to both strands of one allele and give different signals. In those cases all the different signals are combined to form one resulting signal (e.g. 'A' and 'C' 30 together forms 'M', which is the NC-IUB (NC-IUB, 1985) code for this combination).

For each column of the signal matrix the entries for each row

are compared against one another, in other words for each primer the signals produced by the primer for each sequence are compared against each other. A binary matrix is thus generated (18) of the primers with respect to the identity or difference of signals for pairs of sequences. The 5 binary matrix contains non-zero entries where the primer is able to distinguish between a pair of sequences.

The number of pairs of sequences that each primer can distinguish between are counted and a score is allocated to each primer (20) in dependence on the total number of pairs of sequences counted. 10 Thus, the number of non-zero elements for each primer are counted. Primers that are unable to distinguish between any pairs of sequences are rejected (22) and the remaining primers are sorted (24) in order of their score with the primers with the higher scores at the beginning.

A core of primers is created next (26). The primer with the 15 highest score is selected. Where two primers with equal scores exist, the number of positive signals is determined for each and the primer with the greater number of positive signals is chosen. If both primers remain equal, one is then selected arbitrarily over the other. After the main primer has been selected, the first twenty (five times the desired redundancy which is 20 four here) primers giving positive signals for each sequence in turn are selected for the core. All remaining primers are rejected.

A greedy algorithm is then run (28) using the core set of primers to identify the minimum number of primers necessary to distinguish each sequence. As the greedy algorithm is run, primers are added one at 25 a time with each primer being selected in turn in relation to the number of uncovered rows it is capable of covering. When all rows are covered at least four times the reduced set of primers is checked for any sequences that has fewer than four positive signals and extra primers are added as necessary to meet this minimum requirement.

30 A redundancy check is then performed (30) to identify whether any more primers can be removed. During the redundancy check each primer is "tentatively" removed in turn to see whether the remaining

- 15 -

primers meet the minimum requirements.

If not, the next primer is tried. Otherwise the primer is temporarily removed from the set, and the process continues with the next primer in line. This process continues until no more primers can be

5 removed, in which case the last primer to be removed is added back to the set, and the next primer in line tentatively removed and so on. This can be viewed as a depth-first search of a tree where the nodes are combinations of primers, and the number of primers in each node is one less than in a node one level above. The root node thus contains all primers from the

10 greedy algorithm. It has p (the number of primers after the greedy algorithm) primers in it. It also has p child-nodes (because there are p ways in which you can remove one primer from a set of p primers), each with $p-1$ primers. Each of them has $p-1$ children with $p-2$ primers and so on. In this way, all possible combinations of primers in the set fulfilling the

15 requirements are found, and those combinations with the same, least number of primers are saved as the final primer sets.

Instead of applying greedy algorithm to the core set a modified algorithm called CFT may be applied.

20 Lagrangian subgradient

This algorithm consists of three main phases: A subgradient phase where a near-optimal multiplier vector is found, a heuristic phase where a solution to the SCP is found and column-fixing, designed to improve the results of the heuristic phase.

25 In the subgradient phase, a near-optimal multiplier vector u is found using Equation 8. At the beginning, the starting vector u^0 used is defined as

$$u_i^0 = \min_j \frac{c_j}{\sum_{k=1}^m a_{kj}}$$

Equation 9

Later calls use the last vector u before column fixing, and apply a small perturbation before using it as the starting vector. The perturbation is randomly (and uniformly) distributed in the range $\pm 10\%$ for 5 each element. The sequence of multiplier vectors is considered to have converged when the improvement in $L(u)$ in the last 50 iterations is smaller than 0.1%, or when the number of iterations reached $10 \times m$. The factor λ in Equation 8 was set to 0.1 at the beginning, and was updated as follows: Every 20 iterations, the best and worst lower bounds $L(u)$ during those 20 10 iterations are compared to each other. If the difference is larger than 1%, the value of λ is halved. If the difference is less than 0.1%, λ is multiplied with 1.5. In the first call, the upper bound, UB , used is the sum of the costs of the first primers that together cover all rows four times. Otherwise it is the value of the best solution found so far.

15 In the heuristic phase, the last vector from the subgradient phase is used to generate a sequence of multiplier vectors (again using Equation 8), and a feasible solution constructed for each of the multiplier vectors. The procedure used to generate a feasible solution is a variation of the greedy algorithm, where each column is scored according to

20

$$\begin{aligned}\mu_j &= \sum_{i \in R} a_{ij} \\ \gamma_j &= c_j - \sum_{i \in R} u_i^k \\ \sigma_j &= \begin{cases} \gamma_j / \mu_j & \text{if } \gamma_j > 0 \\ \gamma_j \times \mu_j & \text{if } \gamma_j \leq 0 \end{cases}\end{aligned}$$

Equation 10

where R is the set of uncovered rows in each step. The 25 column with the lowest σ_j , i.e. the columns with the best "gain/cost"-ratio, is added in each step to the solution. This continues until no improvements to the best solution (i.e. minimum number of primers) have been made for 50 iterations.

After the heuristic phase column fixing is applied to the solution. Columns that are absolutely necessary in order for a row to be covered (i.e. if there are only e columns covering a row and each row is to be covered e times) are fixed. These fixed columns are then used as a starting point for the greedy algorithm, and the first $\max\{[200/m_i], 1\}$ columns chosen therein are fixed as well.

These three phases are then applied again to the problem, with the condition that the fixed columns must be included in the solution this time. Columns already fixed in a previous round can not be removed from the solution. This goes on until either all rows are covered by the fixed columns, or the cost of the fixed columns is larger than the estimated lower bound for the entire problem or if no new columns were fixed in the last iteration.

When the three phases are done, the problem is refined, in order to improve the solution. Here, each column in the best solution found so far is scored according to

$$\delta_j = \max\{c_j(u^*), 0\} + \sum_{i=1}^m a_{ij} u_i \cdot \frac{K_i - 1}{K_i}$$

Equation 11

where

$$K_i = |S| - \sum_{j=1}^n a_{ij}$$

Equation 12

and S is the set of columns in the solution. The term $u_i(K_i - 1)$ is the contribution of row i to the gap between the estimated lower and upper bound of the problem. This is then split uniformly between all columns in the solution covering that row. Columns with small δ_j (contributing the least to the gap) are then likely to be part of the optimal solution. The p columns with the smallest δ_j are then fixed before the entire

algorithm is applied again to the resulting sub-problem. (Column fixing here has nothing to do with column fixing after the heuristic phase, so columns fixed there need no longer be fixed here). p is the smallest value satisfying

5

$$\frac{|\cup_{j=j_1}^{j_p} I_j|}{e \times m} \geq \pi$$

Equation 13

where $\{j_k\}$ is the set of columns in the solution ordered with ascending δ_j , and I_j is the set of rows covered by column j . π is in the range 0...1 and controls the percentage number of rows removed after fixing. $\pi = 1$ means that no rows will be uncovered, while $\pi = 0$ means that no columns will be fixed before reapplying the algorithm. (Since each row has to be covered multiple times, in this case it is not actually the number of rows but the number of elements covering the rows that are regulated by π). In the beginning, π is set to 0.3 and is multiplied with $\alpha = 1.1$ if the best solution so far was not improved in the last application of the three main phases. If a better solution was found, π was reset to 0.3. Because of the density of the matrices, the number of columns fixed in this step was also set to be at least one more than in the previous iteration (if no improvements were made). Otherwise the same number of columns would be fixed in a number of iterations before the value of π is large enough to allow more columns to be fixed.

The algorithm is iterated until either the value of the best solution is less than the estimated lower bound, all columns in the best solution found so far are already fixed in the refining step or a time limit is exceeded. The time limit in this case was arbitrarily set to as many seconds as there were rows in the problem. However, the time limit is only checked before the refining step. If it is not exceeded, a whole iteration of the algorithm will be executed before another check is done. Here too a

check was done afterwards to see if primers could be removed without breaking any constraints.

With this algorithm no pricing is performed. Pricing is used to update the core problem, exchanging columns between the core problem and columns outside the core. It was not included here since it was argued that since the costs of the columns are all the same, the best columns would be those with the largest number of non-zero elements. These would be the first columns to be added to the core, and the columns not included in the core would most probably not be better than those included.

Also, the pricing step will require some computation which will extend the time required by this algorithm. As is, the computational requirement of this algorithm is several orders of magnitudes higher than for the greedy algorithm. Finally, the main memory available in the computer puts a limit on the how large the problems can be. If pricing was included all data will not fit into the physical memory, forcing the computer to use a swap-file which would increase the computation times considerably.

Using both alternative algorithms described above a minimum number of primers were identified for various sequences. The results are set out below.

It will be apparent that the initial manual rejection of primers, steps (12, 14 and 22) need not be performed and instead the algorithms can be applied to the original complete set of primers. However, the initial rejection of obvious failed primer candidates can significantly reduce the computational time required in the later stages. Similarly, in many cases the final redundancy check (30) need not be performed as in many cases little or no reduction in the number of primers was achieved by this final check.

Furthermore, although in the method described above the primers were initially sorted in order of score, this need not be performed. The algorithms for stripping out redundant primers are capable of operating with any order of primers including a wholly random order. However, slightly better results were obtained when ordering by score was

performed.

Collecting sequences

The HLA-sequences were available internally from

5 Amersham Pharmacia Biotech (release December 1997), and included 91 alleles from HLA-A, 202 HLA-B, 47 HLA-C, 11 HLA-DPA1 (coding for the α -chain), 74 HLA-DPB1 (β -chain), 18 HLA-DQA1, 34 HLA-DQB1, 192 HLA-DR1 and 35 sequences in all of HLA-DR3, -DR4 and -DR5. The length of these sequences range from ~250bp to ~1100bp.

10 The 16S rRNA-sequences were collected from GenBank (Benson et al., 1998), an annotated database of all publicly available DNA sequences. Only a subset of all the available 16S rRNA-sequences were used. The sequences used were all from organisms that could be identified using either the *MicroLog* or the *MicroStation* system from Biolog Inc., or the *API* systems from CounterPart Diagnostics. These systems utilise differences in metabolism in order to identify the organisms, which is the most common way of identifying micro-organisms today. Altogether, 1207 sequences from 523 different organisms were collected from GenBank. 269 of those 523 organisms had only one 16S rRNA sequence among those 1207 sequences. The length of these sequences is between 20 ~1000bp and ~1500bp.

Data set	No. sequences	Mean length of sequences
DPA1	11	517
DPB1	74	288
DQA1	17	616
DQB1	34	490
DRB1	192	324
DRB345	35	400
HLA-A	91	944
HLA-B	200	900
HLA-C	47	1003
16S.rRNA	1207	1452

Table 1: Details about data sets.

The program was written using the Microsoft® Visual C++®, version 5.0 compiler. It was executed on a PC with a Pentium® MMX 233 MHz processor, 64 MB RAM and Windows® 95, unless otherwise indicated. All execution times are for the entire program, including I/O.

As can be seen in Table 2, the binary SCP matrices were quite dense. The density (i.e. the number of non-zero elements in the matrix) usually lies around a few percent, of course depending on the application. A higher density means that fewer columns are needed in order to cover all rows. This is offset in this case by the fact that all rows were required to be covered multiple times. Another consequence of this high density is that the number of primers needed according to the greedy algorithm could be much higher than in the optimal solution. (Recall that the worst case behaviour of the greedy algorithm is a function of the largest column-sum of elements.)

15 **Table 2:** Some details about the binary SCP matrix. Data are calculated for all primers in the primary set.

The program could be considered as consisting of two phases. The first phase involves constructing all primers and finding out what kind of signal they will get for each sequence. The second phase is the optimisation phase, where the SCP is solved. Some details about the first phase can be found in Table 3.

Dataset	DPA1	DPB1	DQA1	DQB1	DRB1	DRB345	HLA-A	HLA-B	HLA-C	16S rRNA
No. rows	55	2701	136	561	18336	595	4095	19900	1081	727821
Density (%)	47.89	20.73	36.31	42.18	24.98	37.70	36.31	32.33	30.41	2.04

25 **Table 3:** Number of primers in different stages of the algorithm and time to get signals for all primers. The number of primers in the core are for homozygotes.

Dataset	DPA1	DPB1	DQA1	DQB1	DRB1	DRB345	HLA-A	HLA-B	HLA-C	16S rRNA
First set	1747	1885	2487	2891	3891	3031	4756	4994	4293	247877
Primary set	1333	1475	2166	2730	3651	3016	3886	4585	3354	247877
Core set	106	321	213	244	385	203	595	750	338	2377
Time (s)	4.67	6.81	11.26	18.51	42.29	14.56	124.74	286.82	61.29	150632

One explanation to this high density is that the sequences in the data sets are quite similar to each other, so that most primers will hybridise to and give signal for more than one sequence (either the same or different signals). This is also indicated in Table 3, where for some data sets there is a noticeable drop from the number of primers in the first set to the number of primers in the primary set. Most of this reduction is due to a primer having the same signal for all sequences, which in turn means that all sequences have a substring that is similar enough for the primer to hybridise to and that the nucleotide after the primer is the same for all sequences. In contrast, the 16S rRNA data set has a much lower density, and no reduction in the primers going from the first set of primers to the primary set. As the sequences in this data set come from organisms which might be only distantly related to each other, there need not be as much similarity between the sequences as there is in the HLA data sets. Another explanation is this: If all k sequences except one give the same signal for a primer, that column in the binary SCP-matrix will have $k-1$ non-zero elements. The density (for that column) will then be $(k-1) / (k(k-1)/2) = 2/k$. In other words, the density will be higher for smaller values of k , and smaller for larger values. This means that it would be "natural" for smaller matrices to have higher densities, and larger matrices to have lower densities.

In the second phase, solving the SCP, a few different approaches were tried. The results, the minimum number of primers needed and the time required to find this number, can be found in Table 4 and Table 5. Even though the worst case behaviour of the greedy algorithm is not so good in this application, the results are not much worse than when using a Lagrangian subgradient (CFT) method. The greedy algorithm typically needs two or three more primers, while the computation times are much lower for the greedy algorithm.

The results show that it is worthwhile to check the results from the greedy algorithm for redundancy. In all cases except one primers could be removed and the resulting primer sets still fulfil all requirements.

This is not true for the CFT algorithm, however, as there is only one instance in which the result could be improved. On the other hand, since there is some randomness in the CFT algorithm (an old multiplier vector is disturbed randomly before being used as a starting vector in the next iteration), the results can differ from one execution of the algorithm to another. Sometimes the results can be improved, and sometimes not (results not shown).

Dataset	DPA1	DPB1	DQA1	DQB1	DRB1	DRB345	HLA-A	HLA-B	HLA-C	16S rRNA
Greedy	11	42	32	31	48	24	73	103	51	210
Time (s)	0.27	1.37	0.61	0.71	11.5	0.66	4.61	31.36	1.15	9921.48*
Final	11	41	30	29	44	21	72	99	47	197^
Total (s)	0.27	1.81	0.72	0.88	30.3	0.71	6.48	85.14	1.76	>300000^

Table 4: No. of primers after the greedy algorithm and time spent by it. Also final nr. of primers after check for redundancy and the total time spent solving the SCP. *Value from a 300MHz Pentium II with 512MB RAM running Windows NT 4.0. ^The computation was halted before completion due to time constraints.

Dataset	DPA1	DPB1	DQA1	DQB1	DRB345	HLA-A	HLA-C
CFT	10	38	26	27	20	69	47
Time (s)	10.22	2748.92	60.80	372.56	427.32	4547.33	1091.37
Final	10	38	26	27	20	69	45
Total (s)	10.22	2749.14	60.86	372.61	427.38	4548.49	1111.70

Table 5: Results using modified algorithm CFT.

One reason CFT is not much better than the greedy algorithm could be that it was designed for other instances of SCP. The SCP arising in this application differ in three aspects from those: A) The density is much higher, B) All rows are to be covered multiple times and C) The costs of all columns are all the same.

A comparison was made between the results from the greedy algorithm and from CFT in Table 6. Most of the primers (70% or more) were chosen by both algorithms, indicating that these primers are likely to

be part of an optimal solution. However, this is only an indication as the only way to prove this is to find an optimal solution. This will require far too much time even for the smallest data set as the problem is NP-hard.

Dataset	DPA1	DPB1	DQA1	DQB1	DRB345	HLA-A	HLA-C
Greedy	11	41	30	29	21	72	47
CFT	10	38	26	27	20	69	48
Same	7	33	22	22	14	62	38
Percent (%)	70.00	86.84	84.62	81.48	70.00	89.86	80.85

5

Table 6: Comparison of primers from the two different algorithms.

Results from combining HLA sequences in order to differentiate between heterozygous individuals can be found in Table 7.

10 CFT was only used for the two smallest data sets due to the time requirements. It performed slightly better than the greedy algorithm on those, but only by one primer on each data set. There are heterozygotes that can not be distinguished from another heterozygote, which can be seen in Table 7. This happens because the combination of two sequences to form

15 one heterozygote could result in exactly the same signal pattern as another combination of homozygotes. In other words, some rows in the signal-matrix will be the same leading to some rows in the binary SCP-matrix not containing any non-zero elements at all. For some of those pairs listed, this is not true, however. They are listed because there were not enough

20 primers that have different signals for these pairs, and so could not meet the requirement of at least four different signals in the signal patterns (Table 8). For the rest, it is simply a limitation of this technique to type HLA-genes. To be able to identify the alleles forming each heterozygote, primers that amplify alleles selectively should be used in the PCR step.

25 This will remove the ambiguities as some heterozygotes simply will be transformed to homozygotes since only one of the alleles in the heterozygote will be amplified and not the other.

Dataset	DPA1	DPB1	DQA1	DQB1	DRB345	HLA-A	HLA-C
Greedy	26	130	51	81	94	172	94
Time (s)	0.99	9229.57	7.41	294.51	453.19	20826.20*	1212.59
CFT	25		50	-	-	-	-
Time (s)	1943.82	-	8427.82	-	-	-	-
Amb. het.	0	16	2	2	6	19	4
Percent (%)	0.00	0.58	1.31	0.34	0.95	0.45	0.35

Table 7: Results from heterozygous pairs. Number of primers needed, the time spent, how many heterozygotes that did not differ by at least four signals from any other heterozygote and the percentage of total number of heterozygotes. *Value from a 300MHz Pentium II with 512MB RAM running Windows NT 4.0.

Unfortunately, it was not possible to obtain any results for 10 heterozygotes for the data sets DRB1 and HLA-B, as these were too large to run on existing machines. A very approximate extrapolation of the primers needed for these data sets suggests that the total number of primers for all HLA sets together would be <1000, which can placed on one chip without problem (one chip can contain up to ~5000 primers). Without 15 the reduction obtained above, at most two genes could be tested on each chip. With the reduction, all nine HLA genes and the 16S rRNA gene can be tested on one chip, and with plenty of room to spare for other genes as well. This makes APEX more versatile, as it allows a family of related genes to be tested using only one chip instead of several.

- 26 -

DPB1	DQA1	HLA-A
Pair 1 DPB1*0501 DPB1*2101 Pair 2 DPB1*2201 DPB1*3601 No. diff. 2	Pair 1 DQA1*0101 DQA1*0104 Pair 2 DQA1*0101 DQA1*0105 No. diff. 3	Pair 1 A*0101 A*2411N Pair 2 A*0104N A*2402 No. diff. 0
Pair 1 DPB1*0501 DPB1*5501 Pair 2 DPB1*3001 DPB1*6301 No. diff. 2	DQB1 Pair 1 DQB1*0604 DQB1*0612 Pair 2 DQB1*0608 DQB1*0609 No. diff. 2	Pair 1 A*0201 A*0205 Pair 2 A*0202 A*0208 No. diff. 1
Pair 1 DPB1*0601 DPB1*3601 Pair 2 DPB1*2001 DPB1*2101 No. diff. 1	DRB345 Pair 1 DRB4*01011 DRB4*01011 Pair 2 DRB4*01011 DRB4*0301N No. diff. 0	Pair 1 A*0201 A*0205 Pair 2 A*0214 A*0222 No. diff. 1
Pair 1 DPB1*0801 DPB1*1401 Pair 2 DPB1*1001 DPB1*5701 No. diff. 0	Pair 1 DRB4*01011 DRB4*0103 Pair 2 DRB4*0103 DRB4*0301N No. diff. 0	Pair 1 A*0201 A*0208 Pair 2 A*0205 A*0220 No. diff. 0
Pair 1 DPB1*0901 DPB1*3001 Pair 2 DPB1*1701 DPB1*5401 No. diff. 0	Pair 1 DRB4*0201N DRB4*0201N Pair 2 DRB4*0201N DRB4*0301N No. diff. 0	Pair 1 A*0201 A*0213 Pair 2 A*0212 A*0226 No. diff. 2
Pair 1 DPB1*0901 DPB1*3601 Pair 2 DPB1*2101 DPB1*3501 No. diff. 0	HLA-C Pair 1 Cw*1203 Cw*1602 Pair 2 Cw*12042 Cw*1601 No. diff. 0	Pair 1 A*0201 A*2406 Pair 2 A*0222 A*2413 No. diff. 0
Pair 1 DPB1*0901 DPB1*4501 Pair 2 DPB1*1001 DPB1*1401 No. diff. 0	Pair 1 Cw*12042 Cw*1502 Pair 2 Cw*1205 Cw*1503 No. diff. 0	Pair 1 A*0202 A*0206 Pair 2 A*0214 A*0222 No. diff. 0
Pair 1 DPB1*3901 DPB1*5301 Pair 2 DPB1*4001 DPB1*4801 No. diff. 0		Pair 1 A*0212 A*2801 Pair 2 A*0222 A*2808 No. diff. 2
		Pair 1 A*2402 A*2502 Pair 2 A*2407 A*2501 No. diff. 0
		Pair 1 A*2402 A*68012 Pair 2 A*2407 A*68031 No. diff. 0
		Pair 1 A*2501 A*88012 Pair 2 A*2502 A*88031 No. diff. 0

Table 8: Heterozygous pairs that do not differ enough in their signal patterns, and how many signals they differ with.

5

The results of this work are summarised in the following

Table 9

Class I	Number of alleles	Primers needed	Class II	Number of alleles	Primers needed
HLA-A	91	172	DPA1	11	26
HLA-B	200	<1000	DPB1	74	130
HLA-C	47	94	DQA1	17	51
			DQB1	34	84
			DRB1	192	<1000
			DRB345	35	94

Table 9. Number of primers needed to discriminate between heterozygote HLA samples.

5

Some sets of primers indicated in Table 9, and also the set indicated for 16S rRNA, are set out in appendix 2.

Primers can be arranged on the surface of a support in such a way that different studied types, genes, alleles, species etc. form easily recognised characters such as figures or letters. These character forming primers can be additional primers of common origin from the gene of interest and be used for validation of the process.

The following demonstration is based on the HLA Class II DQB gene.

15

Experimental

Materials

Amplification:

20 DNA: Four homozygote for DQB cell lines, with alleles 0402, 0301, 06011 and 0201.

Primers: Primer DQB 9246 from Williams *et al.* -96 and DQB 96012 from Amersham Pharmacia Biotech HLA DQB typing kit, covering exon 2,

generating a fragment of 300 base pairs.

Amplification reagents: PCR mix from the Amersham Pharmacia Biotech HLA DQB typing kit, a prototype kit.

All amplifications were spiked with dUTP, to get a final concentration of 100
5 or 200 mM dUTP.

Enzymes for fragmentation of PCR products:

Shrimp alkaline phosphatase (SAP) 1 U/ μ l APB.

Uracil-DNA-glycosylase, (if from PE UDG = UNG) 1 U/ μ l NE Biolabs.

10 SAP will degrade (dephosphorylate) all free dNTPs and UDG will remove all dU from the DNA and after heating the strands will be broken at these points. This step is applicable to any DNA fragment.

15 Primers for spotting:

All 84 primers for the 500 bp fragment were ordered from LTI/GIBCO BRL Custom primers service. All were 25-mers with an amino-activated 5' –end. For primer sequences see appendix 1. Self extended primers were N, A, C, G and T as controls with the following sequences:

20 N: amino TTT AGC CTT AAC GCC T N TGAC GTCA

A, C, G, T: amino TTT AGC CTT AAC GCC T X TGAC GTCA, where X is A, C, G or T.

Extension reagents for the APEX reaction

25 Dyes: Specially synthesised for Baylor by Du Pont and /or APB
Cy2 – ddCTP (equal to fluorescein) 50 μ M
Cy3 – ddATP 50 μ M
Texas Red – ddGTP 50 μ M
Cy5 – ddUTP (often written as T in many of the reactions and
30 results) 50 μ M
10x ThermoSequenase™ DNA polymerase buffer (TS):

260 mM Tris-HCl pH 9.5; 65 mM MgCl₂, ThermoSequenase DNA polymerase (Amersham Pharmacia Biotech) 4 U/μl, if needed dilute with T.S. dilution buffer (=10 mM Tris-HCl pH 8.0; 1 mM β-mercaptoethanol, 0.5% Tween – 20(v/v), 0.5% Nonidet P-40 (v/v). TS was used from a 150 unit stock and diluted 1 μl + 37 μl dilution buffer.

5

Methods

Preparation of glass slides before spotting of primer:

Arrange 25-30 cover slips (24 x 60 mm) in a stainless staining tray.

10

Immerse the tray in glass staining dish with acetone to fully immerse slides.

Place the glass staining dish in sonicator for 10 minutes.

15

Remove the tray from acetone bath, shake off excess of acetone and rinse several times (at least twice) in MilliQ water.

Immerse tray in 100 mM NaOH and sonicate for 10 minutes (a few more minutes, no problem).

Remove the tray and shake off excess of NaOH and rinse several times (at least twice) in MilliQ water.

20

Immerse tray in silane solution and sonicate for 2 minutes.

Wash slides by immersion in 100% EtOH once.

Dry the tray with the slides using nitrogen with a high velocity (without breaking the slides).

25

Cure the slides in a vacuum oven at 100°C over night or until they are used for spotting (at least 20 minutes vacuum is needed).

Spotting of oligos:

All spotting was done with a spotter with 96 parallel capacity.

Each slide was spotted with three replicas of the primers.

30

After spotting the slides were allowed to air dry for 5 to 15 minutes, when dried they were marked. They were stored at room temperature, in a dry place, in the trays until used.

- 30 -

DQB amplification

5 The DQB amplification was done according to the method described by Williams et al. -96 using a 33% dUTP mix. After 40 cycles (95°C, 30 sec.; 55°C, 30 sec.; 72°C, 30 sec.), one microliter of the PCR products was tested on a 1.5% agarose gel, before the fragmentation step.

Williams, Bassinger, Moehlenkamp, Wu, Montoya, Griffith, McAuley, Goldman, Maurer: Strategy for distinguishing a new DQB1 allele (DQB1*0611) from the closely related DQB1*0602 allele Tissue Antigens, 1996, 48:143-147.

10

Fragmentation of PCR products:

Before APEX can be done all DNA fragments must be fragmented so all new fragments can get access to the primer on the chip.

15 Set up:

5 µl DNA from a PCR reaction (1/10 of the PCR reaction)

2 µl SAP (Shrimp alkaline phosphatase) 1U/µl APB

1 µl UDG (Uracil-DNA-glycosylase) 1U/µl NE Biolabs

15 µl water

20 Total: 23 µl

Incubate 37°C for 2 hour.

The samples were frozen and stored until they were used.

Inactivation of enzymes at 100°C for 10 minutes can be done, but not needed since this is the first step in the APEX reaction.

25

Extension method for the APEX reaction

Slide treatment:

30 Start with washing the slides in hot water (90 - 98°C, not boiling) for 2 x 5 minutes in a 50 ml Flacon tube. When the slides are ready, remove them from the tube with a forceps and place them on a dry

- 31 -

heater block at 48°C. The slide(=DNA chip) is now ready for adding the reactions.

APEX reactions set up:

5

23 µl DNA from the fragmentation step.

3 µl 10x TS reaction buffer (the rest of the buffer comes from PCR and UDG cleavage)

17 µl for cover slip method.

10 Heat denature at 100°C for 7 – 10 minutes, target 8 minutes, not longer.

Spin the tube quickly and add quickly

1 µl ThermoSequenase DNA polymerase (4U)

1 µl Dye-mix (50 µM of the four dideoxynucleotides A, C, G, and T, separately dye labelled).

15 Then the reaction mix was physically spread out over the primer array with the tip of a pipette tip. Incubate at 48°C until no trace of solution is seen. This takes about 8 minutes.

Wash with hot water for 2 – 5 minutes, 2 times. Ready to read on detection instrument.

20

Detection

The detection system is a total internal reflection fluorescence (TIRF) system, where microscopic slides are placed on top of a prism with oil on to link a laser beam in to the glass slide. The system has light of five different wave lengths from five different lasers to vary between. In this experiment only four were used. To detect Cy2 a laser with 488 nm was used, for Cy3 a 532 nm, for Cy5 a 635 nm and for Texas Red a 670 nm laser were used. Image related software were based on Image Pro Plus 3.0.

30

Results

Amplification of HLA DQB alleles

The DNA from the four DQB homozygote cell lines were
5 amplified according to the protocol in Williams *et al.* -96 with two different concentrations of dUTP. In addition to this, DNA from six different heterozygotes were amplified. All amplifications worked well and the expected 300 bp fragment were seen from all samples.

10 APEX reaction with DQB chip

Primer chips were washed and fragmented PCR products were incubated on the chip according to the protocol. The image was compared to the expected pattern. The expected pattern was similar to but somewhat different from the recorded pattern, the reason for this is that the 15 set up was planned for a 500 bp fragment, but the actual fragment used was a 300 bp PCR fragment.

Homozygous cell lines results

Figure 4 shows the results from a cell line homozygous for
20 the DQB 0204 allele. The pattern shown in the image is very close or similar to the expected results from exon 2.

In all reaction the control primers worked well and the four dyes were used in the same frequencies. In the case with a 500 bp fragment for DQB typing the primers for allele 0402 were placed in such a 25 way that they formed figures. In Figure 4, panel D, most signals are seen forming a "2" from the 300 bp fragment, and the missing signal will be seen when the large PCR fragment is used. This clearly shows that primers can be placed in a clever way to form figures.

30 Heterozygous results

For the heterozygous test only one of the four dye reactions worked. Some of the expected spots from the heterozygous sample were

not seen, but this is probably due to the fact that no control signals were seen in the lower right hand corner, where the signals were weaker than in other part of the slide.

As this experiment shows, a limited number of primers can be used for HLA typing and if they are placed in a clever way the interpretation of the results is very simple. Both homozygous and heterozygous samples can be correctly analysed with this method.

Continuation

An algorithm was developed in order to select the minimum number of primers needed to identify different genes using APEX. It was applied to the following HLA genes: HLA-A, HLA-B, HLA-C, HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRB1 and HLA-DRB345. It was also applied to the 16S rRNA gene. In the case of HLA-DQB1, the primers have been shown to work as intended. As is, a few assumptions were made (such as how many mismatches to be allowed between the primers and the sample DNA) that need to be tested and possibly refined.

Another improvement that can be made is the following: As is, the program works only with discrete signals, e.g. either there is a signal 'A' or there is not, either there is a signal 'G' or there is not and so on. A more precise approach would be to predict how strong the signals will be for each primer on each sequence. A rough estimate of the signal strength should be possible given some thermodynamic data about the primers, most notably their melting points. With this information, and knowing the concentration of DNA in the sample among other things, the proportion of primers on the chip that will actually react with the sample DNA should be possible to estimate. It would thus allow a rough estimation of what strength the different signals will have. It will not be very precise, and the estimate might possibly be off by a factor 2 or more, but it will still give some information about what signals to expect from the chip.

Given the melting points of the primers, the temperature at which the reaction on the chip is carried out could be optimised as well.

Since the sequences are known, it is possible to estimate the melting point of any primer to any sequence when there are a few mismatches. This could be done for all primers on all sequences, and a range of temperatures calculated. The actual temperature to use could then be chosen so as to be as optimal for as many primers on as many sequences as possible, instead of as now at a standard temperature.

Another possibility would be to try other heuristics to solve the resulting SCP. Even though CFT does give better results than the greedy algorithm, it is not by much. It could be that Lagrangian relaxation methods really are not suitable for unicost problems, but the only way to find out is to try heuristics based on other ideas. It might be possible to reduce the binary SCP-matrix as well, before applying any heuristic on it. Some rows in the matrix could end up the same, in which case one of them could be removed in order to reduce the number of rows and thus speed up computation. No figures of how many rows might be the same exist, but it could be worthwhile examining this possibility to reduce problem size.

The algorithm itself could be improved. The complexity of the redundancy-check phase can be slightly reduced by having a vector consisting of the sums of the rows in each node. For each child-node, the column to be removed is then subtracted from this vector of sums. This operation can be carried out in $O(m)$, and the final complexity will then be $O(m \times N(p, p))$ instead. For the greedy algorithm, another possible improvement is to check the primer set for redundancy each time a primer was added. The complexity for the greedy algorithm will be the same, as the check will take $O(m \times p)$ (i.e. same as each iteration in the greedy algorithm) each time (with the improvement just mentioned). The check could take longer, but that is unlikely as that would imply that one primer could make several other primers redundant. The main advantage is, of course, that no redundancy check with its rather high complexity is needed afterwards.

The most serious problem is the sheer size of the problems. For the 16S rRNA data set, around 300 MB is required just in order to store

all the primers and their signals. Add to that the fact the all primers need to be traversed once for every iteration in the greedy algorithm, and the result is that it will take quite some time as well. This also means that it is not even feasible to use more elaborate algorithms such as the CFT algorithm 5 on the 16S rRNA data set, unless a much more powerful computer is available. On the other hand, algorithm CFT would probably benefit quite a lot from a parallel computer, since much computation could be carried out as vector-operations. It should then be possible to spread out all computations on several processors, thus reducing the time required. It 10 would also reduce the memory requirements on each processor (but then parallel computers tend to have enough memory to store all necessary data for this problem on each processor anyway). Even the greedy algorithm would benefit from a parallel computer, as each processor can be charged with the task of scoring only a subset of primers. It is not as critical in this 15 case, though, since the computation times are not very high when using the greedy algorithm.

As is, this method is only capable of identifying known gene-variants. If applied to a sample with a previously unknown variant, it is very probable that this new variant will be falsely identified as one of the known 20 variants. It would be very advantageous if this method could be augmented in some way to recognise this fact, and give a warning if there could be an unknown variant in the sample. It could be done by giving a warning when the signal pattern gained differs from the signal pattern from any known variants, but this might not be enough. There is no guarantee 25 that the new variant could not differ in some place not affecting any of the existing primers, which would lead to the new variant being indistinguishable from any of the known variants. Some other way is probably needed as well.

APPENDIX 1**Primer sequences for DBQ heterozygote typing**

Primers 'dqb1 -1' to 'dqb1 -8' placed in positions A3-A10.

Primers 'dqb1 -9' to 'dqb1 -18' placed in positions B2-B11.

5 Primers 'dqb1 -19' to 'dqb1 -30' placed in positions C1-C12.

Primers 'dqb1 -31' to 'dqb1 -42' placed in positions D1-D12.

Primers 'dqb1 -43' to 'dqb1 -54' placed in positions E1-E12.

10 Primers 'dqb1 -55' to 'dqb1 -66' placed in positions F1-F12.

Primers 'dqb1 -67' to 'dqb1 -76' placed in positions G2-G11.

Primers 'dqb1 -77' to 'dqb1 -84' placed in positions H3-H10.

dqb1-1 NH2 - TCC ATC ACA GGA GTC AGA AAG GGC T

dqb1-2 NH2 - GTG TGC AGA CAC AAC TAC GAG GTG G

15 dqb1-3 NH2 - GCG GTG ACG CTG CTG GGG CTG CCT G

dqb1-4 NH2 - TAA TGA GGG GGG TGG ACA CAA CGC C

dqb1-5 NH2 - GCG GTG ACG CCG CTG GGG CCG CCT G

dqb1-6 NH2 - GGA CAT CCT GGA GGA GGA CCG GGC G

dqb1-7 NH2 - GTG GTG ACG CCG CTG GGG CCG CCT G

20 dqb1-8 NH2 - TCC GTC AAA GGA GTC AGA AAG GGC T

dqb1-9 NH2 - GAT GTA TCT GGT CAC ACC CCG CAC G

dqb1-10 NH2 - CCG AGT ACT GGA ATA GCC AGA AGG A

dqb1-11 NH2 - GAT GTG TCT GGT CAC ACC CCG CAC G

dqb1-12 NH2 - GGG TGG ACA CAA CGC CGG CTG TCT C

25 dqb1-13 NH2 - GGG TGG ACA CAA CGC CGG TTG TCT C

dqb1-14 NH2 - CTT CTG GCT ATT CCA GTA CTC GGC G

dqb1-15 NH2 - TTC CCG GCG GTG ACG CTG CTG GGG C

dqb1-16 NH2 - GCT TCG ACA GCG ACG TGG GGG TGT A

dqb1-17 NH2 - GCT GTT CCA GTA CTC GGC GCT AGG C

30 dqb1-18 NH2 - CTT CTG GCT GTT CCA GTA CTC GGC G

dqb1-19 NH2 - ACC GTG TCC AAC TCC GCC CGG GTC C

dqb1-20 NH2 - CAC AAC GCC GGT TGT CTC CTC CTG G

dqb1-21 NH2 - CTC CTC CTG GTC ATT CCG AAA CCA C

dqb1-22 NH2 - CCA GGA TCT GGA AAG TCC AGT CAC C

35 dqb1-23 NH2 - GAG CGC GTG CGT CTT GTA ACC AGA T

dqb1-24 NH2 - GAC ATC CTG GAG AGG AAA CGG GCG G

dqb1-25 NH2 - AGA GAC TCT CCC GAG GAT TTC GTG T

dqb1-26 NH2 - TAG TTG TGT CTG CAC ACC CTG TCC A

dqb1-27 NH2 - ACG TAC TCC TCT CGG TTA TAG ATG T

40 dqb1-28 NH2 - GCT TCG ACA GCG ACG TGG AGG TGT A

dqb1-29 NH2 - TCC GTC CCA TTG GTG AAG TAG CAC A

dqb1-30 NH2 - TGA TAA GGC CCA GCC CGA GGA AGA T

dqb1-31 NH2 - GGG TGG ACA CAA CGC CAG TTG TCT C

dqb1-32 NH2 - GGG TGG ACA CAA CGC CAG CTG TCT C

45 dqb1-33 NH2 - GAC AGC GAC GTG GAG GTG TAC CGG G

dqb1-34 NH2 - TCC GTC CCG TTG GTG AAG TAG CAC A

dqb1-35 NH2 - GCA CGA CCT TGC AGC GGC GAC CCC A

dqb1-36 NH2 - GAA CAG CCA GAA GGA AGT CCT GGA G

dqb1-37 NH2 - CTT CTG GCT GTT CCA GTA CTC GGC A

50 dqb1-38 NH2 - AAC GCC AGC TGT CTC TTC CTG GTC A

dqb1-39 NH2 - GAG AGG ACC CGG GCG GAG TTG GAC A

dqb1-40 NH2 - GCA GGC GGC CCC AGC GGC GTC ACC A

dqb1-41 NH2 - GTC GCT GTC GAA GCG CAC GTC CTC C

dqb1-42 NH2 - CTC TGT CCT GGA TGG GGT CGC CGC T

55 dqb1-43 NH2 - ACG GGA CGG AGC GCG TGC GTT ATG T

dqb1-44 NH2 - GAA GTA GCA CAT GCC CTT AAA CTG G

dqb1-45 NH2 - TCG GTG GAC ACC GTA TGC AGA CAC A

dqb1-46 NH2 - GGA M CGT GTA CCA GTT TAA GGG C

dqb1-47 NH2 - ACG TAC TCT TCT CGG TTA TAG ATG T

dqb1-48 NH2 - GAG AGG ACC CGA GCG GAG TTG GAC A
 dqb1-49 NH2 - ACC CCA GCC TCC AGA GCC CCA TCA C
 dqb1-50 NH2 - CAA CGG GAC GGA GCG CGT GCG GGG T
 dqb1-51 NH2 - ACA TCT ATA ACC GAG AGG AGT ACG C
 5 dqb1-52 NH2 - GAA CAG CCA GAA GGA CAT CCT GGA G
 dqb1-53 NH2 - CCT TCT GGC TAT TCC AGT ACT CGG C
 dqb1-54 NH2 - TTA AGG CCA TGT GCT ACT TCA CCA A
 dqb1-55 NH2 - TTC AGA TTG AGC CCG CCA CTC CAC G
 dqb1-56 NH2 - ATC TGG TCA CAA GAC GCA CGC GCT C
 10 dqb1-57 NH2 - AGT AGC ACA GGC CCT TAA ACT GGT A
 dqb1-58 NH2 - ATG TAT CTG GTC ACA CCC CGC ACG A
 dqb1-59 NH2 - ATC TGG TCA CAT AAC GCA CGC GCT C
 dqb1-60 NH2 - ATC AAA GTC CAG TGG M CGG AAT G
 dqb1-61 NH2 - ACG TGG GGG TGT ATC GGG TGG TGA C
 15 dqb1-62 NH2 - ATC AAA GTC CGG TGG M CGG AAT G
 dqb1-63 NH2 - GTA TCT GGT CAC ACC CCG CAC GAG C
 dqb1-64 NH2 - CGC TGT CGA AGC GCA CGT CCT CCT C
 dqb1-65 NH2 - GGA M CGT GTT CCA GTT TAA GGG C
 dqb1-66 NH2 - TGT GGG CTC CAC TCT CCT CTG CAA G
 20 dqb1-67 NH2 - ACG TCC TCC TCT CGG TTA TAG ATG T
 dqb1-68 NH2 - TTG CAG CGG CGA CCC CAT CCA GGA C
 dqb1-69 NH2 - GAA GTA GCA CAG GCC CTT AAA CTG G
 dqb1-70 N H2 - GAA GTA GCA CAT GGC CTT AAA CTG G
 dqb1-71 NH2 - TCG ACA GCG ACG TGG GGG TGT ACC G
 25 dqb1-72 NH2 - TCG ACA GCG ACG TGG GGG AGT TCC G
 dqb1-73 NH2 - TGT GGG CTC CAC TCG CCG CTG CAA G
 dqb1-74 NH2 - CGG CGT CAG GCC GCC CCT CGC GGG T
 dqb1-75 N H2 - TCG ACA GCG ACG TGG AGG TGT ACC G
 dqb1-76 NH2 - GCG TTG GAG GCT TCG TGC TGG GGC T
 30 dqb1-77 NH2 - CGG TGA CCC CGC AGG GGC GGC CTG A
 dqb1-78 NH2 - ATG GGA CGG AGC GCG TGC GTT ATG T
 dqb1-79 NH2 - CGG TGA CGC CGC TGG GGC GGC TTG A
 dqb1-80 NH2 - ACG GGA CGG AGC GCG TGC GTC TTG T
 dqb1-81 NH2 - TGA TAA GGC CAA GCC CAA GGA AGA T
 35 dqb1-82 NH2 - GAG ACT CTC CCG AGG ATT TCG TGT A
 dqb1-83 NH2 - CGT CGC TGT CGA AGC GCA CGT CCT C
 dqb1-84 NH2 - GAC TCT CCC GAG GAT TTC GTG TAC C

40 APPENDIX 2

Homozygotes

(From CFT if available, otherwise greedy algorithm).

DPA1

45 TTTTTTTTTTTGCCAGGGCACAG
 TTTTTTTTTTTTAAGGAAAAGGCTC
 TTTTTTTTTTTGGATCTGGACAA
 TTTTTTTTTTTCTGGCCCAGCTCC
 TTTTTTTTTTTGTACAGACCCA
 50 TTTTTTTTTTTAGGGGACCTGTG
 TTTTTTTTTTTGGCGGACCATGTG
 TTTTTTTTTCTGCTCATCTTCA
 TTTTTTTTTGTCAACTTATGCC
 TTTTTTTTTTCAGGCCGCCAAT

DPB1

|||||||TTCAACCGGGAGGGAG
|||||||TTGGCCTGACGAGGA
|||||||TTAACCTGGAGGGAG
5 |||||TTCCAGTACTCCTC
|||||||TTGCCGTAACTGGT
|||||||TTGGGGCGGCCTGA
|||||||TTGCGCGTACTCCTC
|||||||TTGGACAGGGAGGAA
10 |||||TTACACGGAGGGAGCA
|||||||TTTGCTCCTCTGT
|||||||TTGGCAATGCCCGCT
|||||||TTGGCACTGCCCGCT
|||||||TTAGAGAATTACGTG
15 |||||TTCCAGAGAATTAC
|||||||TTAACTACGAGCTGG
|||||||TTGGTCATGGGCCCG
|||||||TTTGACCCCTGCAGCG
|||||||TTTACACGTAATTCT
20 |||||TTGTAACTGGTACAC
|||||||TTCTGACGAGGGAGTA
|||||||TTTACCTTTCCAG
|||||||TTCTGGAAAAGGTA
|||||||TTGAGAATTACCTT
25 |||||TTGCCTGACGAGGGAG
|||||||TTACTGGTGCACGTA
|||||||TTTCCCTCAGGATGT
|||||||TTTGGGGAGGAGCTG
|||||||TTAGCCAGAAGGACA
30 |||||TTTCAGCCAGAAGGAC
|||||||TTTAGTGCCGGACAGG
|||||||TTTATTGCCGGACAGG
|||||||TTTCTGCAGCGCCGA
|||||||TTTAGAGAATTACCTT
35 |||||TTGGACTCGGGCGTG
|||||||TTTACTACGAGCTGGG
|||||||TTTGCTTCGTGCTGGG
|||||||TTTGTCCCTGGTACAC
|||||||TTTGCCTGCAGGGTC
40 **DQA1**
|||||||TTTACATCCTCATCTG
|||||||TTTACACCCCTCATCTG
|||||||TTCAAGTTTACACCA
|||||||TTTCAGCACAATGTC
45 |||||TTTCCAAGTCTCCCG
|||||||TTTGGGGAGACTTGGG
|||||||TTTAATTCTGGCTGT
|||||||TTTACAATCCCAGGGC
|||||||TTTACAACCCCCAGGGC
50 |||||TTTGTGGGCATTGTGG
|||||||TTTCCAACACCCCTCAT
|||||||TTTGGGCCACAGACAA
|||||||TTTCACTGGGCATTGTG
|||||||TTTGGCCTGGATGAGC
55 |||||TTTGTAGGCTCATCCAGG
|||||||TTTCAACACCCCTCATT
|||||||TTTGTAGCACTGGGGACT
|||||||TTTAAAGGGCCATTGTG

5 TTTAAATTCAATGGGTG
TTTACCATATAAGAGGC
TTTCAACCACAAGAGGC
TTTCACCGTAAGAGGC
TTTCTCCCTTCTG
TTTAACTCTCCTCAG
TTTAAATCTCATCAG
TTTCTCCTCCCTTCTG

DQB1

10 TTTATCTTGAGAGGA
TTTCTCTCCAGGATG
TTTGGTCACCGCCCG
TTTGGGAGTTCCGGGC
TTTCGCTCGGGTCTC
15 TTTCCAGTACTCGCG
TTTCTGGGGCCGCCCTG
TTTATGTCTACACCTG
TTTAAAGGGCTTCTGC
TTTAGCATCACCAAGGA
20 TTTGCAGGAGGAGAC
TTTACCAAGGAGGAGAC
TTTGGTTTCGGAATGA
TTTGGGTGTATCGGGT
TTTGTCGGAAAGGGCT
25 TTTGGTTTCGGAATG
TTTCCAGTACTCGGCA
TTTAGCGCACGATCTC
TTTGTCTCTCCTGGT
TTTCAAGGTCAAGCCGCC
30 TTTGCAGGACTCGTGGG
TTTCAAGGTATAGATGT
TTTGTAAACCAGACAC
TTTGTATGCAGACACA
35 TTTACACCCCCGACG
TTTACACCCCCGACGC

DRB1

40 TTTGCAAGTCTCCTC
TTTCTCCTCCGGT
TTTCCACAACCCGGTA
TTTGGCCAGGTGGACA
TTTGCGGTTCCCTGGAG
TTTCAGCCAGAAGGAC
TTTGACTCGCCTCTGC
45 TTTCCAGGACTCGGC
TTTGAATAAACACTCA
TTTGGAGGGACAGGCG
TTTACGTGGTCGGGTG
TTTACTCCAAGAAC
50 TTTACGGTGTCCACCT
TTTGGAGAGGTTACA
TTTCCAGTACTCGGCA
TTTGGAGTACTCTACG
TTTGTGTAAACCTCTC
55 TTTCGGTGCAGCGCG
TTTGGAGGAGTTCTG
TTTGGAAAGACGAGCG
TTTCAAGGAGGTTGTGG

TTTGTACAGGCGGCCG
TTTCCGTTCAAGGAACC
TTTGGAAATCCTCTTGG
TTTGCACAAAGAAACG
5 TTTTACGTTCTTGGAG
TTTCGGACTCCTCTTG
TTTACGGGTGAGTGT
TTTCCAGGAGGAGTTC
TTTGTATTGTCCACC
10 TTTCTGTAGCGCGCGT
TTTAAGATGCATCTAT
TTTACGTCTGAGTGT
TTTCCAGTACTCAGCA
TTTCTGTAGCGCGCGTA
15 TTATCTCTCCACAAC
TTTGAGCTCCTCTTGG
TTTAACCAGGAGGAGT
TTTAGGGCCCGCCTGT
TTGGAGAGCTTCACA
20 TTGGAGAGATTCA
TTTCAACCGCCCGTA
TTTAACTAACGGGGTTG
TTTCCAGTACTGGGCA

DRB345

25 TTTGTATCTGTCCAGG
TTTGAUTGGGTGGTG
TTTCTGTCAAGCGCA
TTTGTGTAAACCTCTC
TTTCTGTGAAGCTCTC
30 TTTCAACCAAGGCCCGC
TTTGGCCAGGTGGACA
TTTGCCTTCCTGGAG
TTTCAAGCGCGCGT
TTTAACCAAGGAGGAG
35 TTTACGTGGTCGGGTG
TTTAGGGCCCGCCTGT
TTGGGCCCGCCTGTC
TTTAACTAACGGAGTTG
TTGGGCCCGGCTGT
40 TTGACCATGTTCTT
TTCTGTGCAGGAACC
TTGGCCGGCTGTT
TTTACATCTGGAAAGA
TTTCTCACGAGTCCTG

45 **HLA-A**

TTTCACTGTGAGT
TTTAGACGCATATGAC
TTTGGACGCATATGAC
TTTGGTCGCCAGGTCC
50 TTTCCGCAGGCTCTCT
TTTCTCCCTCCACAT
TTTCCGAACCTCGTC
TTTATTTCTCCACATC
TTTGGCGGACATGGCG
55 TTTCCAGAGCGAGGAC
TTTCAACCACATCCG
TTTGGGAGGCCTGCCA
TTTGTATGTGGAGGAG

TTTGGAGGAGGAACAG
TTAGTCATATGCGTC
TTGGTCTGCCGAGC
TTAACCTGCCATGT
5 TTTCCGGACACGGAA
TTCGTCTGGGGGG
TTCCGCTGCCAGGTC
TTATGCGTCTGGGG
TTATGCGTCTGGGG
10 TTGGAGAAGAGATAAC
TTGGGAGCCCCCCA
TTCCGCAGGTTCTCT
TTGCGCAGGTCTCT
TTGGGCGGGCTCTCA
15 TTCCAGGACACGGAG
TTCCGGCAGTGGAGA
TTAGGAGACAGGGAA
TTGTCAATCTGTGAG
TTAGAAGTGGGTGGC
20 TTTCAAGTAGGCTCTC
TTCGGACGCCCAA
TTCAATCTGTGAGT
TTGAAGGCCAGTC
TTCGTCGTAAGCGTC
25 TTTAACCAAGAGCGAGG
TTTGACGGTCATGGC
TTGGACCTGGCGAC
TTGAGAGCCGCCCA
TTCATATTCCGTGT
30 TTTGGGAGACACGGAA
TTGTCCACTCGGTCA
TTCCGTGTCTCCCCG
TTGCTGCCACGTGGG
TTCGAACTGCGTGT
35 TTTGGTAGGCTCTCAA
TTAGGTCCACTCGGT
TTGTCCTGGGGGGGT
TTGCTGCTCCGCCGC
TTGGGGCGCCATGAC
40 TTTGCGCGATCCGAG
TTGCACATGGCAGGT
TTAGGAGAAGAGATA
TTAGGAGCAGAGATA
TTCCACTCCACGCAC
45 TTTCCCGTCCACGCAC
TTTACGTGCCATCCA
TTCCCGGCCGGCAG
TTACGTGCGAGCCAT
TTACGTGGCAGCCAT
50 TTATCCAGAGGGATGT
TTCGAGCTCCGTGTC
TTACCAAGAGCGAGGA
TTATGAACAGCACGC
TTTACACCCCTCCAG
55 TTTCTACGTGGACAAC

HLA-B

TTTTTTTTTTGGATGGCGCCCCG
TTTTTTTTTCGGCTCAGATCTC
TTTTTTTTTCGGGGCGCCGTG
5 TTTTTTTTTCTCCACTGCTCCG
TTTTTTTTGTGTTGGTCTTG
TTTTTTTTGGGTATGACCACT
TTTTTTTTCCAGGTGATGTA
TTTTTTTTGTCTGCTCCGCC
10 TTTTTTTTTGTAGTAGCGGAG
TTTTTTTTGCTCAGGTCTCC
TTTTTTTTACCAACACACAGA
TTTTTTTTCCGTCGTAGGCCT
TTTTTTTTGTGAGCCTGCGGA
15 TTTTTTTTACATCATCCAGAG
TTTTTTTTGGTTCTCTCGTA
TTTTTTTTGATGTGTCTCTC
TTTTTTTTGCGCCATGACCACT
TTTTTTTTGGCGTCCTGGTCA
20 TTTTTTTTAGGAGGGACCTGAG
TTTTTTTTGCGCCAGGCACAG
TTTTTTTTAGGAGGGGGCCGGA
TTTTTTTTCCGCTGCTCCGCC
TTTTTTTTACACCATCCAGAG
25 TTTTTTTTACACAGATCTAC
TTTTTTTTGGGCATGACCACT
TTTTTTTTCACACAGATCTCC
TTTTTTTTGCGAGTGCCTGGA
TTTTTTTTGGTACCCGCGGA
30 TTTTTTTTCCGTGCGTGGAG
TTTTTTTTAGACACAGATCTT
TTTTTTTTCAGCGACGCCACG
TTTTTTTTCGGGCCGGGACAC
TTTTTTTTCCCCTCCCCAATAC
35 TTTTTTTTGGGCATAACCACT
TTTTTTTTGCCCGCTTCATC
TTTTTTTTCAGGAGCGCAGGT
TTTTTTTTCGTCACGCACAG
TTTTTTTTGAGTCCGAGAGAG
40 TTTTTTTTGACACAGATCTCC
TTTTTTTTAACCACTTAGCC
TTTTTTTTAGGCGTGCTGGT
TTTTTTTTGACCCCTGCTCCGC
TTTTTTTTGGGGCTCCGCAGA
45 TTTTTTTTCCGGTCCCCAATAC
TTTTTTTTGCGGGTCACGGCG
TTTTTTTTAGGGCCAGGGCTC
TTTTTTTTATCCTCTGGAGGG
TTTTTTTTGGCAGACGATGTA
50 TTTTTTTTAGGCGGAGCAGGA
TTTTTTTTCAGCTGCTCCGCC
TTTTTTTTATCTGCGGGAGCCA
TTTTTTTTCGGAGCTGTGGTC
TTTTTTTTCGACCAACAGCTCC
55 TTTTTTTTGAAGAGTTCAAGT
TTTTTTTTCATGTCGCAGCCA
TTTTTTTTCTGGGCTGGCTCC
TTTTTTTTCAACACACAGACT
TTTTTTTTGGCGGAGCAGGA

|||||TTTATGACCAGGACG
|||||TTTCCACTGCTCCGCC
|||||TTTATGACCAGGACGC
|||||TTGGAGGGGCCGGAG
5 |||||TTTGCCTGGACGGC
|||||TTTAGATCTGTATCTC
|||||TTTGCCTGGCATGGCG
|||||TTTCCGGGACATGGCG
|||||TTTCCACAGCTGTCCA
10 |||||TTTCGGGACATGGCGG
|||||TTTCCCCTCCACGCAC
|||||TTTGAAGTGGGAGCCG
|||||TTTCCCAATCCACC
|||||TTTCCCACGATGGGGA
15 |||||TTTCCCAGTCCACC
|||||TTTGAGATCTGAGCCG
|||||TTTCCACGCACTCGC
|||||TTTGACAGCGACGCCA
|||||TTTCGCCGCGGACACC
20 |||||TTTGTAGGAGGAAGAG
|||||TTTCTTTCCACCTGA
|||||TTTCACGTCGAGCCA
|||||TTTCAGGTCGAGCCA
|||||TTTCGTAGCCCAC TGC
25 |||||TTTATCCAGGTGATGT
|||||TTTCCCAATCCACCG
|||||TTTGGGCCTTCCTCC
|||||TTTCCCCTTCATCGC
|||||TTTCCCCCTTCATCG
30 |||||TTTCAACACAGACTTAC
|||||TTTAGGACGGTTGGG
|||||TTTCCCCGAACCGTCC
|||||TTTGAGCTTTCCCTCC
|||||TTTGCTCCCGAGAGCA
35 |||||TTTACTCCATGAGGCA
|||||TTTGCTGTGGTGGTGC
|||||TTTGTCCAGAAGGC
|||||TTTGCCCGCGGAGAGGA
|||||TTTGCCCGGGACAAGG
40 |||||TTTCCGCCCTGTCCGC
|||||TTTCGGGTACCAACAG

HLA-C

|||||TTTGTAGCTGGGAGCC
|||||TTTGGTGCAGGGCTCC
45 |||||TTTGGGTGCAGGGCTC
|||||TTTGGAGGCAGGAGCAGC
|||||TTTACGGCGGAGCAGC
|||||TTTGCAGCGGGAGCAGC
|||||TTTAGCGCGCGGAACC
50 |||||TTTCGGCCCAGGTCTC
|||||TTTGGGCTCCCAGCTC
|||||TTTGCAGCGCGGAACCC
|||||TTTACGGCTTCATCT
|||||TTTGGTTGGGGCTCC
55 |||||TTTACTCCACGCACAG
|||||TTTGGGAGCAGGAGGG
|||||TTTGCAGCGCGAACCC
|||||TTTGAGTCTCTCATC
|||||TTTCCTGCAGCCCCCTC

||||||||||||||CCGCCGTGTCGC
||||||||||||||CCGCTGTGTCGC
||||||||||||||CCAGAATATGTA
||||||||||||||CGGGGAGCCCCGC
5 ||||||||||||GCCGTCGTAGCG
||||||||||||||CGCCAGGCACAG
||||||||||||||GAGCCAGGCACAG
||||||||||||||TAGCCGCGCAGG
||||||||||||||GCTGGACGCAGCC
10 ||||||||||||CCAGTGGATGTA
||||||||||||||CACGCACAGGC
||||||||||||||GCCGTCGTCCGCAG
||||||||||||||GAGGGGAGCCCCG
||||||||||||||CGTGTCCCAGGC
15 ||||||||||||GGCATGACCAGTT
||||||||||||||GGTATGACCAGTT
||||||||||||||ACAACCAGGACA
||||||||||||||GAATATGTATGGC
||||||||||||||GACAGCCAGGACA
20 ||||||||||||CTGGCTGTCTGG
||||||||||||||CTCTAGGACAGC
||||||||||||||AGGCCAGGGCTC
||||||||||||||ATAACCAAGTCG
||||||||||||||CATAGGAGGAAGA
25 ||||||||||||GTGGAGACCAGG
||||||||||||||GCTCTTCTCCAG
||||||||||||||GAAGAATGGGAAG
||||||||||||||GCAGAAACTGCG

16S rRNA

30 ||||||||||||AGCCGCCTGCGT
||||||||||||||GCCGCAAGGCTG
||||||||||||||GAAGTGCCTGTA
||||||||||||||AGACTGCCGCTGA
||||||||||||||ATTGGAAATA
35 ||||||||||||GCACCCCTTGT
||||||||||||||CGCGAGGTTGAGC
||||||||||||||ACCCCCCATTGT
||||||||||||||CATTGATACTGG
||||||||||||||GTGTGCCTAACAC
40 ||||||||||||ACGACTAACCC
||||||||||||||CCCGGCCTTGTA
||||||||||||||GGGCAAATGGAG
||||||||||||||GATTTGATCCTGG
||||||||||||||GACTCCCAGG
45 ||||||||||||GAAGTCGTAGCAA
||||||||||||||CGCTGCAGAGATG
||||||||||||||ACCCCTACCTACT
||||||||||||||GAGGACCTTCGGG
||||||||||||||AAGGGCCATTACC
50 ||||||||||||GATAAACGCTGGC
||||||||||||||GACTAGCTACTCC
||||||||||||||ACATCCGGTGT
||||||||||||||ATCGCAGGCCCTG
||||||||||||||ACCAAGTCGCT
55 ||||||||||||CCCTCCTTCGG
||||||||||||||AACGCTGGC
||||||||||||||CGAAACCGCAAGG
||||||||||||||GCAAGCGTCCTCC
||||||||||||||ACCAAGGACGTTT

5 TTTCTAATACCCGGAG
 TTTTTTTTTACTTTCACTGGGG
 TTTTTTTTTCTGCGTGAAGTCG
 TTTTTTTTTAATAGCCCACCAA
 TTTTTTTTTAACGGAAACGGGG
 TTTTTTTTTGATTGCACCTCTG
 TTTTTTTTTAGCCTGGGGAG
 TTTTTTTTTCGCCGCATGGCTG
 TTTTTTTTTGCTATAAGGGCAT
10 TTTTTTTTTACACACATCTCTG
 TTTTTTTTGTTACCGCGAGGA
 TTTTTTTTGGCTTCAGAGAT
 TTTTTTTTCGCTGCTTCGCTG
 TTTTTTTTAGCGCTACCTTG
15 TTTTTTTTGCACCACCTGTCA
 TTTTTTTTGAGTTAAACCT
 TTTTTTTTCTAATACGGGATA
 TTTTTTTTAGGAGAAAGCTTG
 TTTTTTTTAAGAGATTAGC
20 TTTTTTTTGTAGCATTCTGAT
 TTTTTTTTAGGCTTCCCCCA
 TTTTTTTTAGAAGTAGCTTCG
 TTTTTTTTCGCGTATCATCG
 TTTTTTTTCAAGAGATTAGC
25 TTTTTTTTCCGAAAGCGTGG
 TTTTTTTTACAACCCGAAGC
 TTTTTTTTGTCATGGCTCAG
 TTTTTTTTCGTAGGCTTGGTG
 TTTTTTTTGTGGAATTCCACG
30 TTTTTTTTACGGTTCCCGAAG
 TTTTTTTTAACTCGAGTGCCT
 TTTTTTTTGATGTGCTATT
 TTTTTTTTAAGCAGGGAGGAA
 TTTTTTTTCTGCTGCAGTGAA
35 TTTTTTTTGGGGATTAGCTC
 TTTTTTTTCCCTTGATACTGG
 TTTTTTTTGGACGCTAGCGGC
 TTTTTTTTGTTACTACCCAC
 TTTTTTTTCGCGATCTCTAGC
40 TTTTTTTTAGGCCGTTCCC
 TTTTTTTTACGCGTTGCATCG
 TTTTTTTTGCCCCGTCAAGCCA
 TTTTTTTTAGTCCCCGCCATT
 TTTTTTTTCTAGCCGTAAGGG
45 TTTTTTTTGTCCCTCGGGGG
 TTTTTTTTAAACCAACTCCCAT
 TTTTTTTTACTGTGGGTATA
 TTTTTTTTCTGAAAGATGGCG
 TTTTTTTTCGAAAGCCAGGGG
50 TTTTTTTTGTCCCGGAATTCTG
 TTTTTTTTCAGAAGTGGGTAG
 TTTTTTTTCAGTCCTCATGG
 TTTTTTTTGAAAAGAAGCTTGC
 TTTTTTTTGACCAACCTGTAC
55 TTTTTTTTGGAACACTGCAT
 TTTTTTTTACAGTTCCCGAAG
 TTTTTTTTCTCATATCTCTAC
 TTTTTTTTCAGTGAGGAAG
 TTTTTTTTACTGTGAGGAAGG
60 TTTTTTTTCCCAGCCCGTAAG

|||||||TCGTAGCCTTGGTG
|||||||TTATGATGCGTAGCC
|||||||TTAGGCAGTGGCTCA
|||||||TCAGGACTAACCC
5 |||||GGCCAGGCCGTAAC
|||||||CCAACCTCGTGC
|||||||GAAGCGTGTGTGA
|||||||CTCCCCGAAGGT
|||||||ATGGGAGTTGTT
10 |||||GTGTGCCGTTACC
|||||||TAGCAGTGAGGAAT
|||||||GCCCGGGTTAACT
|||||||GCACCGGCAGTCA
|||||||GGACCTTCCTCTC
15 |||||TACCTAGGTGGGAT
|||||||AATAGCTAATACC
|||||||GCCATATCTCTAC
|||||||GCCGGTGGGTAA
|||||||ACCCCACCTTCG
20 |||||CAAGGCCTGGGAA
|||||||CAACCTGGTGGC
|||||||TAGTCATCCAGT
|||||||GGCTGCTGCCCTCC
|||||||CCCAGAGCTAAC
25 |||||GAAAGCTTGATCC
|||||||AACACGCTGGCAA
|||||||GAGCTTGCTCCCC
|||||||ATTAGTTGAGCA
|||||||CGACTTAGGCTCA
30 |||||GATGTGCTATT
|||||||CTTAGGTGCCAGC
|||||||GGCTACAGATCGT
|||||||AACTGCGTGCAT
|||||||GCGATTACGTCAA
35 |||||GGACGTTGGCGGC
|||||||GGTGGAGCATGT
|||||||ATAAACCATGCGG
|||||||AAGAAGTGGGTAG
|||||||ACAAGCTAATCC
40 |||||CCATGGTTGAC
|||||||AGTAAC TGCCGGT
|||||||CAAAAGGGGGCGT
|||||||GGCGCTTGCCTC
|||||||GCTACCTACGTGC
45 |||||GCGAGGTGGAGC
|||||||CGCGAGGTGGAGC
|||||||GCTACCTACTTCT
|||||||AACACACATAAA
|||||||GTTGTGAAATGT
50 |||||CGTAAAACCTAAA
|||||||CAAGGGGCAAGT
|||||||CCAACCTTGCGG
|||||||GGAGGAACGTGGG
|||||||ATAAGCCTCTCAG
55 |||||ATGCTAATCCCA
|||||||GATGCTAATCCCA
|||||||GCCAGTGTTCGTC
|||||||TAAAGGTGGGGA
|||||||AACACACCGCC
60 |||||CCAAGGCGGTGAT

TTTTTTTTTTGCTACGGCTAACT
TTTTTTTTAGTCGAGCACTCT
TTTTTTTTAAGGGTAGCTAAT
TTTTTTTTGTACAGTACGAG
5 TTTTTTTTGAAAGCACTTTA
TTTTTTTTGGCGCAAGGCTTA
TTTTTTTTGCCTAGGTGGGAT
TTTTTTTTGTCCCCCACGTTCC
10 TTTTTTTGGCCACAAGGGGA
TTTTTTTTCTAGCTGTAGGGGA
TTTTTTTTGTGGGCAGCAAGC
TTTTTTTCGAAAAGATTAAA
TTTTTTGGAGTATGGTCGC
15 TTTTTTTCGAGATGTGAAAG
TTTTTTGGGCAGGCTAGAG
TTTTTTTACCTCCTGAGCCA
TTTTTTTCCACCGCTACAC
TTTTTTTCAGTCTTGCG
20 TTTTTTTCTTGACGGGGCGT
TTTTTTTACGGTAAAAGATG
TTTTTTTCAACCCTTGC GG
TTTTTTTAACCAGAAAGCC
TTTTTTCAACCAGAAAGCC
TTTTTTGTGTCAAAGGCAG
25 TTTTTTTAAGTCCGGATTG
TTTTTTTGCACATGCTGAT
TTTTTTTATCAGCCTGCCGC
TTTTTTTGTGGTAGGGTAA
TTTTTTTGTGGTAGGGTAA
30 TTTTTTTCAACTCAT AAGGG
TTTTTTTCACTGCTTAAA
TTTTTCCGCAGTCCCACC
TTTTTCTAGTCATAAGGG
TTTTTCACTGATTTGACG
35 TTTTTGGCCACACAGGGGA
TTTTTCCCCCATTGT
TTTTTGACCAAGAAAGGG
TTTTTACACTGGGGGATA
TTTTTCAGCCGCCCTCG
40 TTTTTGTGCCAGCTCGT
TTTTTCTCATATGAATTG
TTTTTGTAAAGGGAGCG
TTTTTCTGAAAGGGAGCG
TTTTTGGCGGCTCCCTCC
45 TTTTTCAGATGTTCTCC
TTTTTGTCTCACGACACG
TTTTTCAGCCGCCCTACG
TTTTTGTGCTAATACC
TTTTTCTTGGAACTGCAT
50 TTTTTTAGTACTCACCGT
TTTTTATTGCTCCATCAG
TTTTTGATCCTGAGCCA
TTTTTAGCAAGTAGAACG
TTTTTGCAAGTAGAACG
55 TTTTTGATAACCGCAAGG
TTTTTGCAGCGTTTCC
TTTTTGAATAACCTCCCTT
TTTTTACAGAGCTTACA
TTTTTGTCCCTCGGGAG
60 TTTTTAGGCGGCTTGCTG

Heterozygotes

From CFT if available, otherwise greedy algorithm.

5 DPA1

10 TTTTTTTTTTGTGCCAGGGCACAG
TTTTTTTTTCTGTTGTTCTATG
TTTTTTTTTAAGGAAAAGGCTC
TTTTTTTTTATGAAGATGAGCA
TTTTTTTTTCACCCCTCAGTGAC
TTTTTTTTTGTCAACTTATGCC
TTTTTTTTTGCAGGAAGAGGGCT
TTTTTTTTTGTACAGACGC
15 TTTTTTTTTCGGTCTCCTTCTT
TTTTTTTTTGCATGGGGAGCC
TTTTTTTTTGGATCTGGATAA
TTTTTTTTTGATGAAGATGAG
TTTTTTTTTGTACAGAC
20 TTTTTTTTTCGTGTACAGAC
TTTTTTTTTCTCAGGCCGCCAA
TTTTTTTTTCTCAGGCCACCAA
TTTTTTTTTATGTGGATCTGGA
TTTTTTTTTACACTCAGGCCGC
25 TTTTTTTTTCACACTCAGGCCG
TTTTTTTTTCAGGCCACCAAAC
TTTTTTTTTCTGCTGTACAAAC
TTTTTTTTTAGAACATCTCATC
TTTTTTTTTAGAACTGCTCATC
30 TTTTTTTTTGAATTGATGA
TTTTTTTTTGAGTTGATGA

DPB1

35 TTTTTTTTTCAACCGGGAGGAG
TTTTTTTTTCAACCTGGAGGAG
TTTTTTTTTCTGCTGGAGGAG
TTTTTTTTTGCTGGGGGTCA
TTTTTTTTTGGCCTGACGAGGA
40 TTTTTTTTTTAACTACGAGCTGG
TTTTTTTTTCCAGAGAAATTAC
TTTTTTTTTGCCTGTACTGGT
TTTTTTTTTCCAGTACTCCCTC
TTTTTTTTTACTGCCCCGACAGG
45 TTTTTTTTTTACCCCCCAGCAGG
TTTTTTTTTAGAGAATTACGTG
TTTTTTTTTCCAGTACTCCGC
TTTTTTTTTGCATTCCCTGCCGT
TTTTTTTTTCGGGAGGAGCTCG
50 TTTTTTTTTTCAGCCAGAAGGAC
TTTTTTTTTATTGCCGGACAGG
TTTTTTTTTCTGCAGCGCCGAG
TTTTTTTTTGCAGCGTACTCCCTC
TTTTTTTTTACAGAATTACCTT
55 TTTTTTTTTTAAGTGTACCAG
TTTTTTTTTATCCTGGAGGGAGA
TTTTTTTTTGGTCATGGGCCCG
TTTTTTTTTGGGAGGAGTACGC
TTTTTTTTTGGGGCGGC!CTGA

TTTTTTTTTTTAAAAGGTAATTCT
TTTTTTTTTTCTGCCGTAACTGG
TTTTTTTTTTGTGTCTGCATA
TTTTTTTTTTGGCTGTTCCAGTA
5 TTTTTTTTTTGTCCTGGTACAC
TTTTTTTTTCTGCAGCGCCGA
TTTTTTTTTCTTGGAGGGGGA
TTTTTTTTTGAGGTCCCTCTGG
TTTTTTTTCAACC CGGCAGGAG
10 TTTTTTTTTGTGTCTGCATAC
TTTTTTTTTCGGGAGC.AGTTCG
TTTTTTTTTGACCCCTGC-AGCG
TTTTTTTTTCAGAGAATTACCT
TTTTTTTTGGGTAGAAATCC
15 TTTTTTTTTTACGTGCACCAAG
TTTTTTTTTCGCTGCAGGGTCA
TTTTTTTTTAGCCAGAAGGACA
TTTTTTTTGTTCCAGTAGTCC
TTTTTTTTGGCTGCTGCGGA
20 TTTTTTTTTGCAGCGCCGAGG
TTTTTTTTTACTACGAGCTGGT
TTTTTTTTCTGGGGCGGCCCTG
TTTTTTTTACAGCGACGTGGG
TTTTTTTTGCCGGACAGGAT
25 TTTTTTTTTCTGCCGTCCCTGG
TTTTTTTTCATGGGCCCCGACC
TTTTTTTTGTCCCATTAAACG
TTTTTTTTGTAACTGGTACAC
TTTTTTTTAAGGACCTCCTGG
30 TTTTTTTTTCTCCTGGAGGAGA
TTTTTTTTGAGAATTACGTGT
TTTTTTTTCTGATGAGGTGT
TTTTTTTTCACAGGAGGAGCA
TTTTTTTTGCCGTCCCTGGT
35 TTTTTTTTTGGGAGGAGTCGC
TTTTTTTTGGACAGGAGGAA
TTTTTTTTACCTGCAGCGTC
TTTTTTTTCCGCCCGGAACTC
TTTTTTTTGCTGCAGGGTCAC
40 TTTTTTTTTACAGGACTATCCA
TTTTTTTTGCGTACTCCTGCC
TTTTTTTTCCGTAACTGGTGC
TTTTTTTTGCAGGAATGCTAC
TTTTTTTTCCAGGCAGCATT
45 TTTTTTTTTAACCGGGAGGAG
TTTTTTTTGGCCTC.AGGCGGA
TTTTTTTTACTACGAGCTGGG
TTTTTTTTATGAGGTGTACTG
TTTTTTTTATACATCTACAAAC
50 TTTTTTTTTAACTGGTACACT
TTTTTTTTACGTAATTCTCT
TTTTTTTTAGCATT CCTGCCG
TTTTTTTTACTGGTACACTTA
TTTTTTTTGGCAATGCCCGCT
55 TTTTTTTTTGCTCGTGCTGGG
TTTTTTTTCGCCCGGAACCTCT
TTTTTTTTACAGGACTGTCCA
TTTTTTTTCCCTCCAGGAGGT
TTTTTTTTCTCTGGCTGTT
60 TTTTTTTTTGTTCCAGTACTCC

5 TTTTTTTTTGCCTGCAGGGTC
 TTTTTTTTTAACCTGGAGGAGA
 TTTTTTTTCTGCCGTAAAC
 TTTTTTTTACGCTGCAGGGTC
 TTTTTTTTCCACAGAATTACC
 TTTTTTTTCCAGAGAATTACG
 TTTTTTTTCGCCGAGTCAGC
 TTTTTTTTAACAGGCAGGAGT
 TTTTTTTTCTCCAGGATGT
10 TTTTTTTTAACCGGCAGGAGT
 TTTTTTTTCTCAGAGAATTAA
 TTTTTTTTGTTCCAGTACACC
 TTTTTTTTCTCTGTAGGAGA
 TTTTTTTTACCTTTCCAG
15 TTTTTTTTGGAGGGAGTCGTG
 TTTTTTTTGAGGAGCTCGTGC
 TTTTTTTTGCCGTAACTGGTG
 TTTTTTTTGCCGCTCCTCCT
 TTTTTTTTCGTCCCTGGAAAA
20 TTTTTTTTGCCGTCCTGGAA
 TTTTTTTTCCCTCCAAGAAG
 TTTTTTTTGCTGCCTGGTAG
 TTTTTTTTCCAGTAGTCCTC
 TTTTTTTTATTCCCTGCCGTA
25 TTTTTTCTGGAAAAGGTA
 TTTTTTCTGCCCTGGTACA
 TTTTTTCTCCTCCAGGAAG
 TTTTTTCTGATTCTGCC
 TTTTTTATCTCCCTGCTGG
30 TTTTTTGAAGGACAACCTG
 TTTTTTCGTGCACCAAGTTA
 TTTTTTCGGACAGGGTATG
 TTTTTTCGGACAGGATATG
 TTTTTTGCACTCGGCGCTG
35 TTTTTACACGTAATTCTC
 TTTTTCGTAACTGGTACA
 TTTTTAATGACCCCCCAG
 TTTTTCTCTCCAGGAAG
 TTTTTCAGCGACGTGGGA
40 TTTTTCTCTGCCGGTTGT
 TTTTTGAGGGACATCCTG
 TTTTTGAGGGACCTCCTG
 TTTTTGTTCCAGTACAC
 TTTTTCAGAAGGACAACC
45 TTTTTGCCTGATGAGGTG

DQA1

50 TTTTTTTCAAAAGAGGCAAC
 TTTTTTTCTATAAGAGGCAAC
 TTTTTTTGAACAC.AGGCAAC
 TTTTTTTACATCCTCATCTG
 TTTTTTTGAGTGCCCATTGC
 TTTTTTTCAGCCACAATGTC
55 TTTTTACAATCCCAGGGC
 TTTTTACAACCCCAGGGC
 TTTTTGTGGGCATTGTGG
 TTTTTATGGGCATTGTGG
 TTTTTCCAACACCCCTCAT
60 TTTTTGTAGACTGTGGCTG

TTTTTTTTTTTCCAACATCCTCAT
TTTTTTTTTTGGCCCACAGACAA
TTTTTTTTTTCATGGGCATTGTG
TTTTTTTTTTAACATCCTCATCT
5 TTTTTTTTTCAACACCCCTCATT
TTTTTTTTTTGACTGTGGTCTGC
TTTTTTTTTTAGCACTGGGGACT
TTTTTTTTTTCTTAGATTTGACC
TTTTTTTTTTAGATTTGACC
10 TTTTTTTTTTCGATGTTCAAGTT
TTTTTTTTTTCAATCCCAGGGCG
TTTTTTTTTTCTCGGATGATGA
TTTTTTTTTTCCACACATAGAACT
TTTTTTTTTTAAATTCATGGGTG
15 TTTTTTTTTTCAGCCACAATGCC
TTTTTTTTTTCACCATATAAGAGGC
TTTTTTTTTTCTCCCTTCTG
TTTTTTTTTTAACTCTCCTCAG
TTTTTTTTTTAAATCTCATCAG
20 TTTTTTTTTCTCCTCCCTTCTG
TTTTTTTTTTGTCAGCCACAATG
TTTTTTTTTCATTCCCTTCTTC
TTTTTTTTTTCTTCCTCCCTTCT
TTTTTTTTTTATAACTCTCCTCA
25 TTTTTTTTTTGAGGCTCATCCAG
TTTTTTTTTC,AGGCTTGTCAG
TTTTTTTTTATGTTGACCACAG
TTTTTTTTTAGTGCCCAACCACA
TTTTTTTTGAACATCCTGATT
30 TTTTTTTTTGGACCTGGAGAAG
TTTTTTTTTCCCTCTGGCCAGT
TTTTTTTTTCCCTCTGGGR-AGT
TTTTTTTTTACACCGTAAGA
TTTTTTTTTAGAAGATTTGACC
35 TTTTTTTTTGAACTGGCCAGAG
TTTTTTTTTGCTACAACCTCTAC
TTTTTTTTTCAAGTCTTACGGTC
TTTTTTTTTCAAGTCTTATGGTC
40 **DQB1**
TTTTTTTTTTTATCTGCAGAGGA
TTTTTTTTTTGGCTGGGTGCTC
TTTTTTTTTTGGTCACCGCCCCG
45 TTTTTTTTTCTGGGGCCGCTG
TTTTTTTTCTCGGCGCTAGGC
TTTTTTTTGTATCTGGTCACA
TTTTTTTTAACTACGAGGTGG
TTTTTTTTCCAGTACTCGGCG
50 TTTTTTTTTCGGTTATAGATGT
TTTTTTTTGCAAGTCTGGAG
TTTTTTTTGGACACAAACGCC
TTTTTTTTCTGGGGCTGCCCTG
TTTTTTTTGGCCTTAAACTGG
55 TTTTTTTTTGTGTCTGCATAC
TTTTTTTTGTCGGAAAGGGCT
TTTTTTTTGGGTGTATCGGGT
TTTTTTTTCCAGTACTCGGCA
TTTTTTTTGTAGACATCTCCA
60 TTTTTTTTTAGGAAACGGGCGG

||||||||||||CACACCCGCACG
||||||||||||CCGCTCGGGTCC
||||||||||||AGCATCACCAAGGA
5 ||||||||||||CCAGTTAAGGGC
||||||||||||TAGGCCACAAGGA
||||||||||||GTATGCAGACACA
||||||||||||CCAGTACTCGGC
||||||||||||AGCCACGATCTC
10 ||||||||||||GGACATCCTGGAG
||||||||||||GGGGCTGCCTGA
||||||||||||GTCAAGAAAGGGCT
||||||||||||CAGGAGCCCCCTTC
||||||||||||GTCTCTTCTGG
15 ||||||||||||ACACCCCGCACGC
||||||||||||GGTTTCGGAATG
||||||||||||AACGGGACAGAGC
||||||||||||GCTGGGGCCGCCT
||||||||||||GAGGATTCGTGT
||||||||||||GAGAGGAGTACGC
20 ||||||||||||CACATCAAAGTCC
||||||||||||GCCAGGAGGAGAC
||||||||||||GTACTCGGCGGCA
||||||||||||TCGCCAGTTGTCTC
25 ||||||||||||AGGGGGTGGAACA
||||||||||||AGATGTATCTGGT
||||||||||||GGGGGAGTTCCG
||||||||||||GTCTCCTCTGG
||||||||||||CACACTCTGTCCA
30 ||||||||||||GGAATGATCAGGA
||||||||||||ATGGGTCGCCGC
||||||||||||CAGATCAAAGTCC
||||||||||||AACGGGACCGAGC
||||||||||||AGGAGTACGTGCG
||||||||||||ATGTGACCAGATA
35 ||||||||||||AGGGCGGGCTGT
||||||||||||CGCCGGTTGTCTC
||||||||||||GTAACCAGACAC
||||||||||||GTGAAGTAGCACA
||||||||||||AGCGCGACCCCA
40 ||||||||||||CACACCCGTCCA
||||||||||||GTGTGACCAGATA
||||||||||||GGACCTTCCAGA
||||||||||||ATCGGGTGGTGAC
||||||||||||GTTTAAGGGCTG
45 ||||||||||||GAAGTAGCACAG
||||||||||||GCTCCAACGGTA
||||||||||||CCTAAACTGGTA
||||||||||||AGGAGGACGTGCG
||||||||||||CGTGCTGGGGCT
50 ||||||||||||CGCTGCTGGGGCT
||||||||||||CCAAGGAAGATCA
||||||||||||ACCGCGCGGTGAC
||||||||||||GCCCTTAAACTGG
||||||||||||GGTCACACCCCG
55 ||||||||||||GGGAGTTCCGGGC
||||||||||||AGGAGGAGACAAAC
||||||||||||GGGTGGACACAAAC
||||||||||||CTGCTCGGTGAC
||||||||||||GGGGCGGCTTGA
60 ||||||||||||GCGCACGTCTCC

TTTTTTTTTTTAGGATTCGTGA
TTTTTTTTTTGCCTTAAACTGGA

DRB345

5 TTTTTTTTTTGTACCTGGACAGA
TTTTTTTTTTTGTTCTGGAGAGA
TTTTTTTTTTTACACTCATAACTTA
TTTTTTTTTTTACACTCAGACTTA
10 TTTTTTTTTTTCCTGGAGCAGGC
TTTTTTTTTTTCGAAGCGCGCGT
TTTTTTTTTTTAATCTGCACAGAG
TTTTTTTTTTAGGGCCCGCCTGT
TTTTTTTTTTTAGGACACTCTGGA
15 TTTTTTTTTTGTGTAAACCTCTC
TTTTTTTTTTCTGTCGAAGCGCA
TTTTTTTTTTGGGGCCGGGCGTGT
TTTTTTTTTTCTTCCAGGATGT
TTTTTTTTTTTAACTACGGAGTTG
20 TTTTTTTTTTCAAGAAACATGGT
TTTTTTTTTTTAACCAGGAGGAG
TTTTTTTTTTGAAGCTCTCCAC
TTTTTTTTTTGGGGCGGCCTGTC
TTTTTTTTTTGCAGCGCGCGTGT
25 TTTTTTTTTTCTTGGAGCTG
TTTTTTTTTTCTCTTCCCTGGC
TTTTTTTTTTTAACTACGGGGTTG
TTTTTTTTTTGTATCTGATCAGG
TTTTTTTTTTGGCCAGGTGGACA
30 TTTTTTTTTTGCCCCCAGCTCCGT
TTTTTTTTTTGGTTCTGGAGAG
TTTTTTTTTTGTCGAAGCGCAGC
TTTTTTTTTTGTGTCTGCAGTAG
TTTTTTTTTTGCTCCACTGGCA
35 TTTTTTTTTTTACGGGGTTGGT
TTTTTTTTTTTCGGTTCTGCACA
TTTTTTTTTTTCCAGTACTCGC
TTTTTTTTTTGTCCACCTCGC
TTTTTTTTTTCTTCCCTGGCCGT
40 TTTTTTTTTTGGTGTCCACCAGG
TTTTTTTTTTTACTCCGTAGTTGT
TTTTTTTTTTTCACTCAGACTTAC
TTTTTTTTTTTGATGCTAGAAACA
TTTTTTTTTTGTGGAATGGAGAG
45 TTTTTTTTTTTAACCAAGAGGGAG
TTTTTTTTTTGTTCGGAAATGGC
TTTTTTTTTTGTATCTGCAGTAG
TTTTTTTTTTTACCTCCCTGGTCTG
TTTTTTTTTTTAGCCAACAGGACT
50 TTTTTTTTTTGCGGTTCTGCAG
TTTTTTTTTTCGCGCCGCGGTGG
TTTTTTTTTTGTAAACCTCTCCA
TTTTTTTTTTCTGATCAGGCTCC
TTTTTTTTTTTCCAGGACTCGGC
55 TTTTTTTTTTTAACCATTACAGA
TTTTTTTTTTTCGGGCCCTGGTGG
TTTTTTTTTTGTTCGGAAACGGC
TTTTTTTTTTGCAGGCCCGCCTGT
TTTTTTTTTTTCCCTGGAAGACAC
60 TTTTTTTTTTGCCGGGTGGACAA

T T T T T T T T T C T G C T C C A G G A T G
T T T T T T T T C A A C T A C T G C A G A
T T T T T T T T G T A C C T G G A G A G A
T T T T T T T T A C C T C T C C A C T C C
5 T T T T T T T T G T G A A G C T C T C C A
T T T T T T T T C C G C G G C G C G C G T
T T T T T T T T C T G A T C A G G T T C C
T T T T T T T T A A T G G G A C G G A G C
T T T T T T T T A T G G A A G T A T C T
10 T T T T T T T T C T G C A G T A G G T G
T T T T T T T T C G G G C C G C G G T G G
T T T T T T T T C T G T G C A G G A A C C
T T T T T T T T C C A A G A G G A G G A C
T T T T T T T T C A A T T A C T G C A G A
15 T T T T T T T T C A C C T A C T G C A G A
T T T T T T T T C T G C C T G G A T A G A
T T T T T T T T G T A A T T G T C C A C C
T T T T T T T T C A C C A G G G C C C G C
T T T T T T T T G C G G T A C C T G G A
20 T T T T T T T T C C T G C A G C A C C A C
T T T T T T T T G C G G C G C G C T G T
T T T T T T T T C C A G G A C T C G G C A
T T T T T T T T G A C A C A A C T A C G G
T T T T T T T T G A T A C A A C T A C G G
25 T T T T T T T T A C T C A G A C T T A C A
T T T T T T T T G A G A C T T A C A C A
T T T T T T T T A C G G G G T T G T G G
T T T T T T T T G T A G T T G T C C A C C
T T T T T T T T A A C C A G G A G G A G T
30 T T T T T T T T A A C C A A G A G G A G T
T T T T T T T T C C A C A G C C C C G T
T T T T T T T T C A G C C A G A A G G A C
T T T T T T T T G G A G G A G T T C C T G
T T T T T T T T G A A C T C C T C T G G
35 T T T T T T T T A A C C A C T C A C A G A
T T T T T T T T G G C C G G G C T G T T C
T T T T T T T T C T C A C G A G T C C T G
T T T T T T T T G T C G A A G C G C A A G
T T T T T T T T C C T C C T G G T C T G
40 HLA-A
T T T T T T T T C A G T C T G T G A G T
T T T T T T T T C C G C A G G C T C T C T
45 T T T T T T T T A T G A G G T A T T T C T
T T T T T T T T G G A C A T G G A G G T G
T T T T T T T T C - A G G T A G G C T C T C
T T T T T T T T A C T C T T G G G G G C
T T T T T T T T G G T C G C C A G G T C C
50 T T T T T T T T G G G A G C C C G C C C A
T T T T T T T T C C G C T G C T C C G C C
T T T T T T T T G A A G G C C C A G T C
T T T T T T T T G C A G C C A T A C A T C
T T T T T T T T C C A C T C C A C G C A C
55 T T T T T T T T C A C G T C G C A G C C A
T T T T T T T T G G T C T G C C C G A G C
T T T T T T T T C A G G T A G A C T C T C
T T T T T T T T G G G A G A C A C G G A A
T T T T T T T T C C C G T C C A C G C A C
60 T T T T T T T T G T C C A C T C G G T C A

|||||||||||||ATCCAGAGGATGT
||||||||||||CGCGATCCGCAGG
||||||||||||CCGGGACACGGAA
||||||||||||GGAGGAGGAACAG
5 |||||||||||TAAGTGAAGGCCA
||||||||||||GGGCTGGGGAG
||||||||||||CAGACTAACCGAG
||||||||||||GTCCTGGGGGGT
||||||||||||CGTCGTAAGCGTC
10 |||||||||||TAGGTCCACTCGGT
||||||||||||GGTAGGCTCTCAA
||||||||||||CGCGATCCGCAG
||||||||||||GTGTCCTGGGTCT
||||||||||||ATCC.AGATAATGT
15 |||||||||||CCGTCGTAGGCAGT
||||||||||||CATATTCCGTGT
||||||||||||CGGACCCCCCCC
||||||||||||GCCGCATGGACCG
||||||||||||GCTGCTCCGCC
20 |||||||||||AGCGCAGGTCC
||||||||||||TACCTGGATGGC
||||||||||||GGTATTCTTCAC
||||||||||||ATATGAAGGC
||||||||||||CCGTGTCTCCCG
25 |||||||||||CCGGCAGTGGAGA
||||||||||||CGGACGCC
||||||||||||CCGTGAGGCGGAG
||||||||||||AGGAGACAGGGAA
||||||||||||AGAGCGAGGACGG
30 |||||||||||GCACATGGCAGGT
||||||||||||CAGCTGCTCCGCC
||||||||||||ATGAACAGCACGC
||||||||||||CCCGCCGGCAG
||||||||||||GCAGCCTGAGAGT
35 |||||||||||GACGGTCATGGC
||||||||||||CCGTCGTAAGCGT
||||||||||||GAGTATTGGGACC
||||||||||||CTGGCCTGGTTCT
||||||||||||ACCTCATGGAGTG
40 |||||||||||AGCCGCCATGTCC
||||||||||||CACGTGCCATCCA
||||||||||||GGTCCCCAGGTT
||||||||||||AGGAGAAGACATA
||||||||||||CTGCTGCTCCGCC
45 |||||||||||GACCCAGACCA
||||||||||||CGGGCGGAGCAGT
||||||||||||AGGTTCGCTCGGT
||||||||||||CATATGCGTCTG
||||||||||||CGTCCTGGGGGG
50 |||||||||||GCACGTGCGTGG
||||||||||||GGTATTCTACAC
||||||||||||AGGAGCAGAGATA
||||||||||||CCGAACCCCTCGT
||||||||||||GCCACATGGGCC
55 |||||||||||AGCAGGAGGAGCC
||||||||||||ATCCAGATGATGT
||||||||||||GGATGGGGAGCAC
||||||||||||GC.ACTGGCGCTTC
||||||||||||AGCTTGTAAAGTG
60 |||||||||||GATAATGTATGGC

||||||| |||||CACACCCCTCCAG
||||||| |||||CTACGTGGACAAC
||||||| |||||CGAGCGAACCTGG
||||||| |||||CGAGAC,AGCCTGC
5 ||||| |||||GGGCTACGTGGAC
||||||| |||||ACCACCACTACGC
||||||| |||||GAGGATGTATGGC
||||||| |||||GATCTCAGCCGCC
||||||| |||||GATCTGAGCTGCC
10 ||||| |||||GATGATGTATGGC
||||||| |||||TACACCTGGAGAAC
||||||| |||||GATGTATGGCTGC
||||||| |||||CCGCAGGTTCTC
||||||| |||||GAGCAGAGATAAA
15 ||||| |||||GGGCTGGGAAGAC
||||||| |||||GATGGGCAGGACT
||||||| |||||CACT|||CCCTGT
||||||| |||||CCCACGATGTGGA
||||||| |||||TAGTCATATGCGTT
20 ||||| |||||GGCGGACATGGCG
||||||| |||||GCTCCGCCTCACG
||||||| |||||CGTCGTAAGCGTT
||||||| |||||GATC,ATGTTGGC
25 ||||| |||||GCTCCTCCTGCTC
||||||| |||||ACTCACCGAGTGG
||||||| |||||AGTCATATGTGTC
||||||| |||||GGTCTGAGCTGCC
||||||| |||||CCCACTTGCGCT
30 ||||| |||||GCCCACTCACAGA
||||||| |||||GGCTCAC.ATCAC
||||||| |||||GCTTTGGACCGC
||||||| |||||GAGAGCCTGCGGA
||||||| |||||GGAACACACGGAA
35 ||||| |||||CGGAACACACCGA
||||||| |||||CGTAAGCGTCCCTG
||||||| |||||GCCGGTGCCTGGA
||||||| |||||GCCGCATGGGCCG
||||||| |||||CCAGAGCGAGGAC
40 ||||| |||||CCCAACGGGCCGC
||||||| |||||CGAGTGCCTGGAG
||||||| |||||GCGAACCTGGGA
||||||| |||||CGGGTACCAAGCGG
||||||| |||||GAAGCGGGGCTC
45 ||||| |||||GGCGGCCCGTTGG
||||||| |||||CTGGGTC,AGGGC
||||||| |||||GCCTCATGGGCCG
||||||| |||||CCATCCCCGTGCC
||||||| |||||AGCTCAGACACCAC
50 ||||| |||||GTCGTAAGCGTCC
||||||| |||||CCCGGCCGGGA
||||||| |||||GGTCCAATACTC
||||||| |||||CGTCCAATACTC
||||||| |||||GTTCTCACACCAT
55 ||||| |||||CCTCTGGATGGT
||||||| |||||CCCACTTGCT
||||||| |||||CCTGACCCAGACC
||||||| |||||GAGAGCCCGCC
||||||| |||||GAGTGCCTGGAGT
60 ||||| |||||ACATCATCTGGA

TTTTTTTTTTTGTATCCGCAGGGTTC
TTTTTTTTTTTTTAGAGCAGGAGAG
TTTTTTTTTTTCTGGCAGCGGGAA
TTTTTTTTTTTCATGGAGTGAGA
5 TTTTTTTTTTCCGGCCGCGGGAA
TTTTTTTTTTTCCAGGACACGGAG
TTTTTTTTTTTCCGGGACACGGAG
TTTTTTTTTTTGCAGCCACACATC
TTTTTTTTTTTGGATGGTGTGAGA
10 TTTTTTTTTTAACATCATCTGGA
TTTTTTTTTTTCCCTCCACAT
TTTTTTTTTTGGCGGGAGCAGT
TTTTTTTTTTGCAGGGGATGGA
TTTTTTTTTTCGCAGGAAGCGCC
15 TTTTTTTTTTTGGCCGTATGGCG
TTTTTTTTTTTATGCGTCTGGGG
TTTTTTTTTTTATGCGTCTGGGG
TTTTTTTTTTTCCCTGTCTCC
TTTTTTTTTTTCAGGGTGGCCTC
20 TTTTTTTTTTTGAGGGAGAACAGC
TTTTTTTTTTTGCAGGGTCGCC
TTTTTTTTTTTCAGCCAAACATCC
TTTTTTTTTTTACTCTGGAAGGT
TTTTTTTTTTTCCCTCTGGACGGT
25 TTTTTTTTTTTGGAGAAGAGATAAC
TTTTTTTTTTTATTCCGTGTCTCC
TTTTTTTTTTTCAATCTGTGAGT
TTTTTTTTTTTGGCCCCTGGCG
TTTTTTTTTTTCGGCGGACATGGC
30 TTTTTTTTTTTTACAAGCTGTGAG
TTTTTTTTTTTCGAACCTGGTGTGTC
TTTTTTTTTTTCGAGCTCCGTGTC
TTTTTTTTTTTACTCCACGCACCG
TTTTTTTTTTCTACGTGGACGAC
35 HLA-C
TTTTTTTTTTTGTAGCTGGGAGCC
TTTTTTTTTTTATCACAAACAGCCA
40 TTTTTTTTTTTAGGCTCTCCGCTC
TTTTTTTTTTGGAGTGGGAGCAG
TTTTTTTTTTCACACCCCTCCAG
TTTTTTTTTTTACTCCACGCACAG
TTTTTTTTTTTGCCTCGTAGGCG
45 TTTTTTTTTTCGCGCAGAACCCCC
TTTTTTTTTTTAGTAGCCGCGCAG
TTTTTTTTTTGGAGCGGAC,AGCC
TTTTTTTTTTCAGGTAGGCTCTC
TTTTTTTTTTGGTTCTGGGGCTCC
50 TTTTTTTTTTGCCCCAAGCCCTC
TTTTTTTTTTGGGCATGACCACT
TTTTTTTTTTGCGGCTCCGCGGC
TTTTTTTTTTCCAGTGGATGTA
TTTTTTTTTTGGCATGACCACTT
55 TTTTTTTTTCTCACTCGGTCA
TTTTTTTTCAAGCCCTCCCTCC
TTTTTTTTTAGTTCCGC.AGG
TTTTTTTTTCAGGTGGCAGGCCA
TTTTTTTTCACTGCGATGAAG
60 TTTTTTTTTGGTATGACCACTT

TTTTTTTTTTTACAGCCAGGCCAG
TTTTTTTTTTTGAGGCGGAGCAGC
TTTTTTTTTTTGGTTGTAGTAGC
TTTTTTTTTTTACCTGCAGAACT
5 TTTTTTTTTTTTGGGCCAGGTCTC
TTTTTTTTTTTGCTGGACGCAGCC
TTTTTTTTTTTCAGGTTCCGCAGG
TTTTTTTTTTTCCGCCAGGCACAG
TTTTTTTTTTTCTCTACACATC
10 TTTTTTTTTTTTACGGCGGAGCAGC
TTTTTTTTTTTAGCGCGCGGAACC
TTTTTTTTTTTCACTCGGTCA
TTTTTTTTTTTACGCCGCGAGTCC
TTTTTTTTTTTGGAGCAGGA,GGG
15 TTTTTTTTTTTGGGTATGACCA
TTTTTTTTTTTATACCTGGAGAAC
TTTTTTTTTTTGGGTTCGGGGCTC
TTTTTTTTTTTGACCGCTAGGACA
TTTTTTTTTTTATCTGAGGCCCTG
20 TTTTTTTTTTTCGCGGAGAGCCCC
TTTTTTTTTTTCTGGCGCTTGT
TTTTTTTTTTTCTGCGGAAACTA
TTTTTTTTTTTAGCGTCTCCTTCC
TTTTTTTTTTTGGCGCCCCGAAC
25 TTTTTTTTTTTATGATGTGAGACC
TTTTTTTTTTCTCGGTGTCCCTGG
TTTTTTTTTTTGTAGTAGCCCGGT
TTTTTTTTTTTAGGATGTGAGACC
TTTTTTTTTTTGGTAGGCTCTCTG
30 TTTTTTTTTTTAGCGTCTTCTTCC
TTTTTTTTTTTCA TAGGAGGAAGA
TTTTTTTTTTTGACAACCAGGACA
TTTTTTTTTTTGC CGCGGGGAGCC
TTTTTTTTTTTGGTGAGGGGCTCT
35 TTTTTTTTTTCGAGGGGCTGCCA
TTTTTTTTTTGGGTATAACCA
TTTTTTTTTTTCCAGAAATATGTA
TTTTTTTTTTTGGGTGCAGGGCTC
TTTTTTTTTTTCGCGCGGAACCCC
40 TTTTTTTTTTTTAGTAGCCCGCTA
TTTTTTTTTTTAGCTGCTCTCAGG
TTTTTTTTTTTACCGCACGA
TTTTTTTTTTTCCGCAGGCTCACT
TTTTTTTTTTTGGTGAGACCCG
45 TTTTTTTTTTTGGAGCCCCGAAC
TTTTTTTTTTTAGCCGCGGGAGCC
TTTTTTTTTTTACTGCACGA
TTTTTTTTTTTCCGCACGA
TTTTTTTTTTTGGTGCA
50 TTTTTTTTTTTGCAGCAGGAGC.AG
TTTTTTTTTTTGAGTCTCTCATC
TTTTTTTTTTTCCGCCGTGTCCGC
TTTTTTTTTTTCCACGCACAGGC
TTTTTTTTTTTACTCGGTCA
55 TTTTTTTTTTCACACC.ATCCAGA
TTTTTTTTTTTCA CACACCCTCCAGA
TTTTTTTTTTTGCAGCAGGATGAG
TTTTTTTTTTTCAGCCACCA
TTTTTTTTTTTCTGTGGCTGGCCT
60 TTTTTTTTTTTTACGGCGGAGCAG

TTTTTTTTTTCTCACACCATCCA
TTTTTTTTGGCGGCGGAGCAG
TTTTTTTTCTGAGCCGCCGT
TTTTTTTTGGCGGAGCAGCAG
5 TTTTTTTTCCGCTGCAGACAC
TTTTTTTTATAACCAGTTCG
TTTTTTTTCACATCCTCCAGA
TTTTTTTTCCGTGTCCGGCGC
TTTTTTTTTCGTGGACGACACA
10 TTTTTTTTCCGCTGTGTCCGC
TTTTTTTTGAAGAATGGGAAG

CLAIMS

5 1. A method of identifying a set of extendible primers for use in
the identification, typing or classification of a nucleic acid of known
sequence having known polymorphisms wherein:

10 i) all possible nucleotide sequences of a chosen length of the
nucleic acid are identified and their corresponding extendible primers,
ii) at least one extendible primer is removed from the set
wherein the at least one primer removed identifies a segment of the nucleic
acid identified by at least one other primer.

15 2. The method of claim 1, wherein between steps i) and ii):
ia) potential extensions for each primer are identified with
respect to each nucleotide sequence,
ib) for each extendible primer the identified potential extensions
are compared to determine which pairs of sequences can be discriminated
by the primer.

20 3. The method of claim 1 or claim 2, wherein a matrix of primers
and pairs of primer extensions is prepared in binary form and is subjected
to analysis by a set covering problem (SCP) algorithm.

25 4. The method of claim 3, wherein a greedy algorithm is used.

5. The method of claim 3, wherein a CFT algorithm is used
which involves a Lagrangian relaxation heuristic.

30 6. The method of any one of claims 3 to 5, wherein a set of core
primers is selected as a base for analysis by the SCP algorithm.

7. The method of any one of claims 3 to 6, wherein the set of extendible primers identified by the SCP algorithm is subjected to a redundancy check.

5 8. A set of extendible primers, for use in the identification, typing or classification of a nucleic acid of known sequences having known polymorphisms, identified by the method of any one of claims 1 to 7.

9. The set of extendible primers of claim 8, in the form of an
10 array.

10. The set of extendible primers of claim 8 or claim 9, for use in the identification, classification or typing of an organism, allele or gene selected from class 1 HLA, class 2 HLA and 16S rRNA.

15 11. The set of extendible primers of any one of claims 8 to 10, wherein the primers are arrayed on a surface of a support in such a way that recognisable patterns are formed with different types or alleles.

20 12. A set of extendible primers, for use in the identification, typing or classification of a human leucocyte antigen (HLA) gene as indicated, the set comprising about the number of primers indicated and being capable of distinguishing about the number of alleles indicated:

	HLA gene	Number of Alleles	Number of Primers
Class I	HLA-A	91	172
	HLA-B	200	<1000
	HLA-C	47	94
Class II	DPA-1	11	26
	DPB-1	74	130
	DQA-1	17	130
	DQB-1	34	84
	DRB-1	192	<1000
	DRB345	35	94

13. A set of extendible primers, for use in the identification, typing or classification of 16S rRNA, wherein set comprises about 210 primers and is capable of distinguishing at least about 1207 different sequences.

5 14. The set of extendible primers of claim 12 or claim 13, wherein the primers have variable segments substantially as set out in appendix 1 or appendix 2.

10 15. A method of identification, typing or classification of a nucleic acid of known sequence having known polymorphisms, by the use of the set of extendible primers as claimed in any one of claims 8 to 14, which method comprises applying the nucleic acid or fragments thereof to the set of extendible primers under hybridisation conditions, and effecting template-directed chain extension of extendible primers that have formed hybrids.

15 16. The method of claim 15, wherein the set of extendible primers is provided in the form of an array, and template-directed chain extension is effected using labelled chain-terminating nucleotide analogues.

20 17. The method of claim 16, wherein template-directed chain extension is effected using four different fluorescently-labelled chain terminating nucleotide analogues, and the results are analysed by total internal reflection fluorescence or confocal microscopy.

25 18. The method of any one of claims 15 to 17, wherein the nucleic acid is a PCR amplimer.

30 19. The method of any one of claims 15 to 18, wherein the nucleic acid is HLA Class 1 or HLA Class 2 or 16S rRNA or a PCR amplimer thereof.

20. The method of any one of claims 15 to 19, wherein a dUTP/uracil-DNA-glycosylase system is used to break the nucleic acid into fragments.

5 21. A kit for use in the identification, typing or characterisation of a nucleic acid of known sequence having known polymorphisms, comprising the set of extendible primers as claimed in any one of claims 8 to 14.

10 22. The kit of claim 21, comprising also a pair of primers for effecting PCR amplification of the nucleic acid.

15 23. An array of sets of extendible primers as claimed in any one of claims 8 to 14, for the simultaneous identification typing or classification of two or more different HLA genes.

20 24. A computer readable storage medium having a program recorded thereon, wherein the program consists of instructional steps for identifying a set of extendible primers for use in the identification, typing or classification of a nucleic acid of known sequence having known polymorphisms, the steps comprising:

i) identifying all possible nucleotide sequences of a chosen length of the nucleic acid and their corresponding extendible primers.

ii) removing at least one extendible primer from the set wherein the at least one primer removed identifies a segment of the nucleic acid identified by at least one other primer.

25. Computer readable program implement consisting of instructional steps for identifying a set of extendible primers for use in the identification, typing or classification of a nucleic acid of known sequence having known polymorphisms, the steps comprising:

5 i) identifying all possible nucleotide sequences of a chosen length of the nucleic acid and their corresponding extendible primers.

 ii) removing at least one extendible primer from the set wherein the at least one primer removed identifies a segment of the nucleic acid identified by at least one other primer.